Robust Statistics using Stata

UK17 Stata Users Meeting

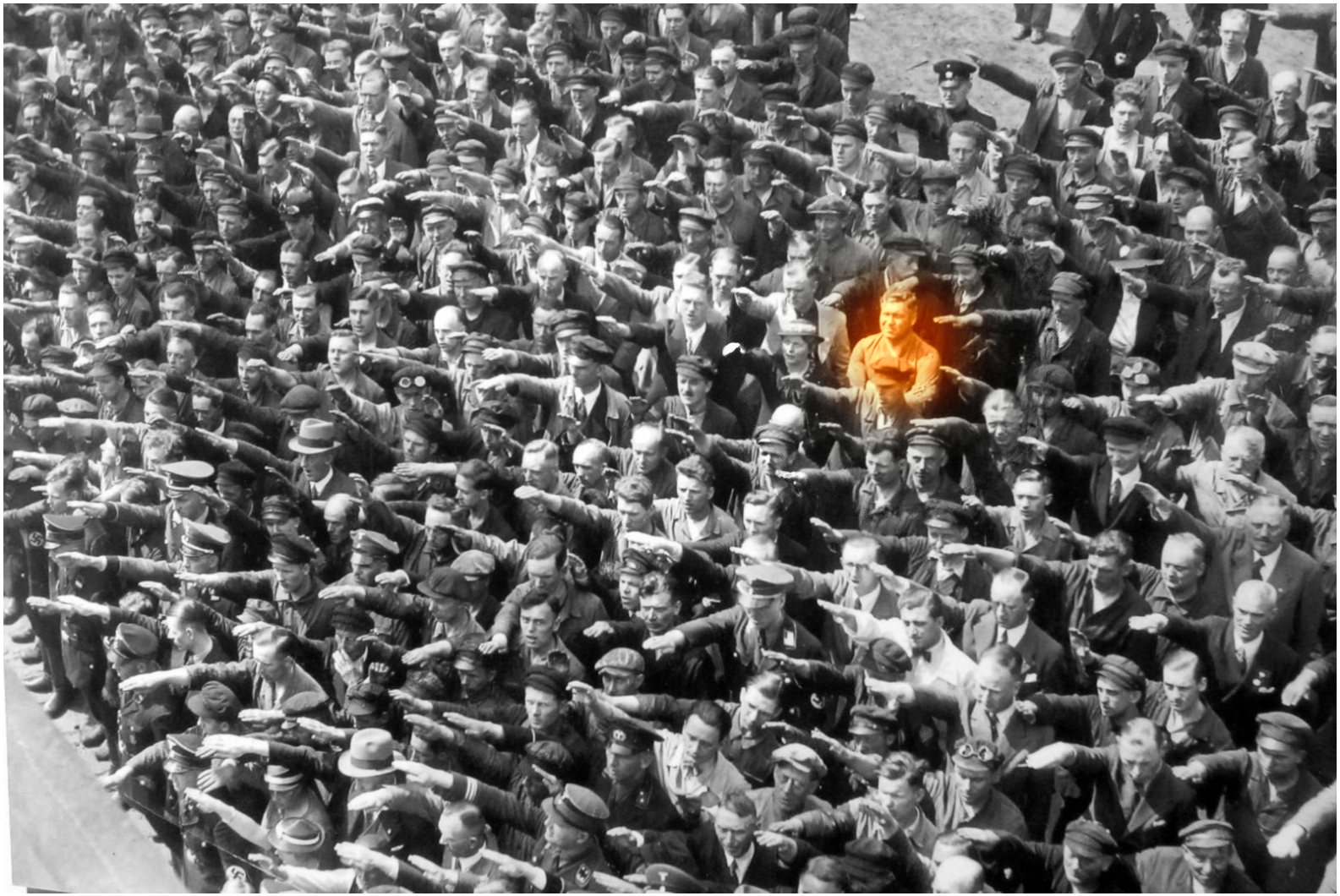
Vincenzo Verardi

Fnrs, UNamur, ULB

September 2017

Outliers do matter and are not always bad

August Landmesser

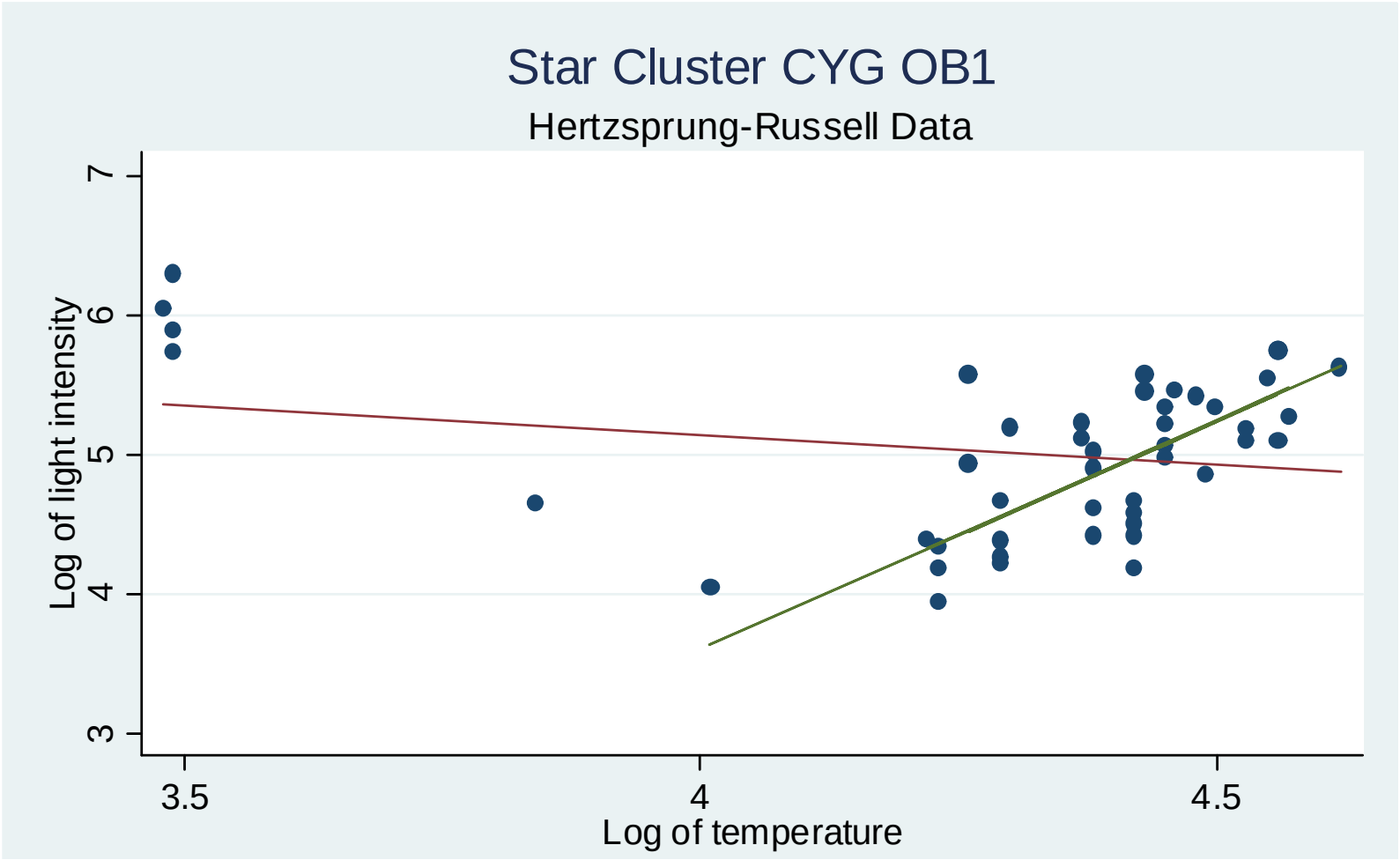


Outliers do matter and are not always bad

Structure of the presentation

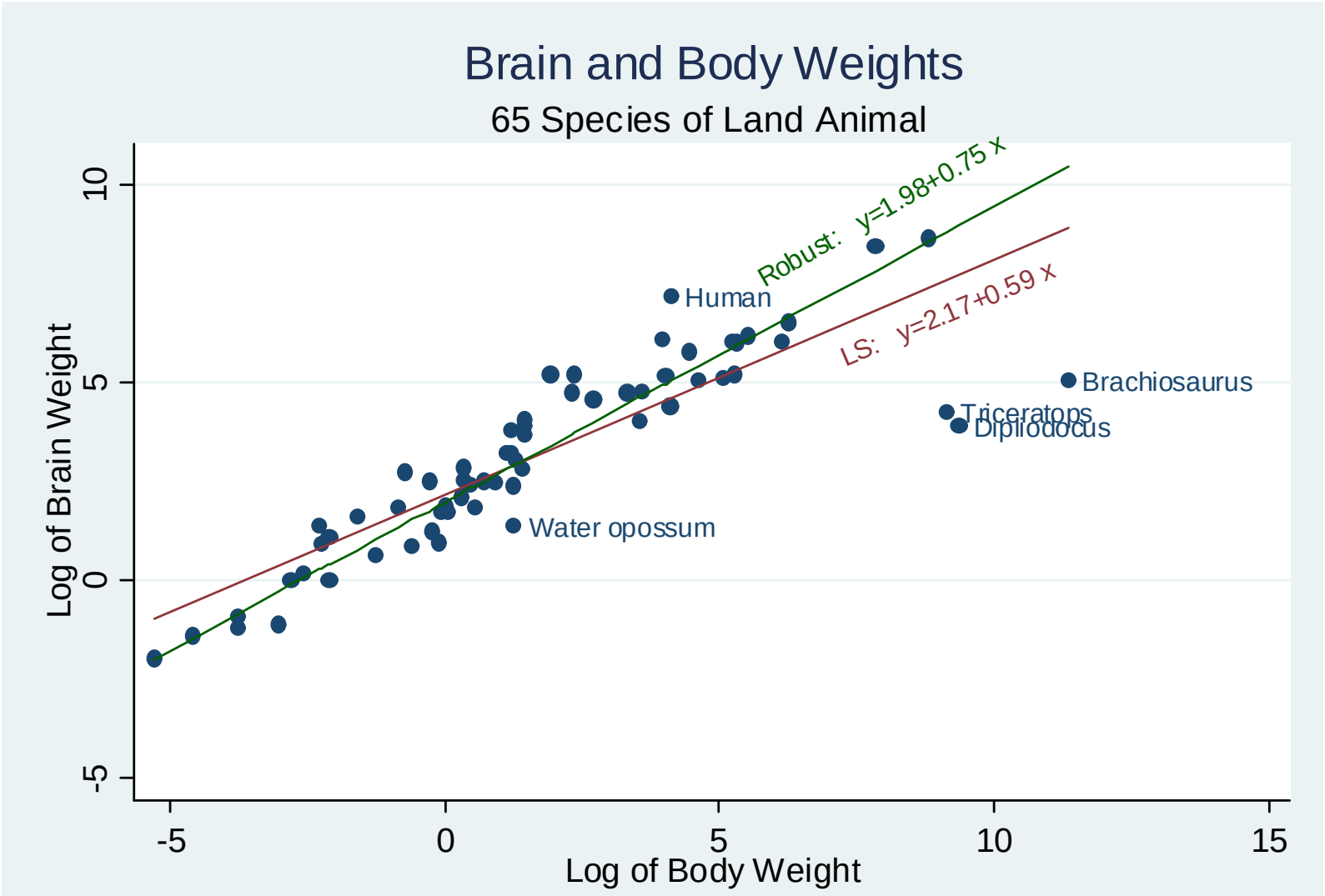
- Introduction
- Descriptive Satistics
- Univariate outliers identification
- Regression models
- Multivariate analysis
- Multivariate outlier identification
- Robust logit
- Conclusion

Outliers do matter and are not necessarily coding errors



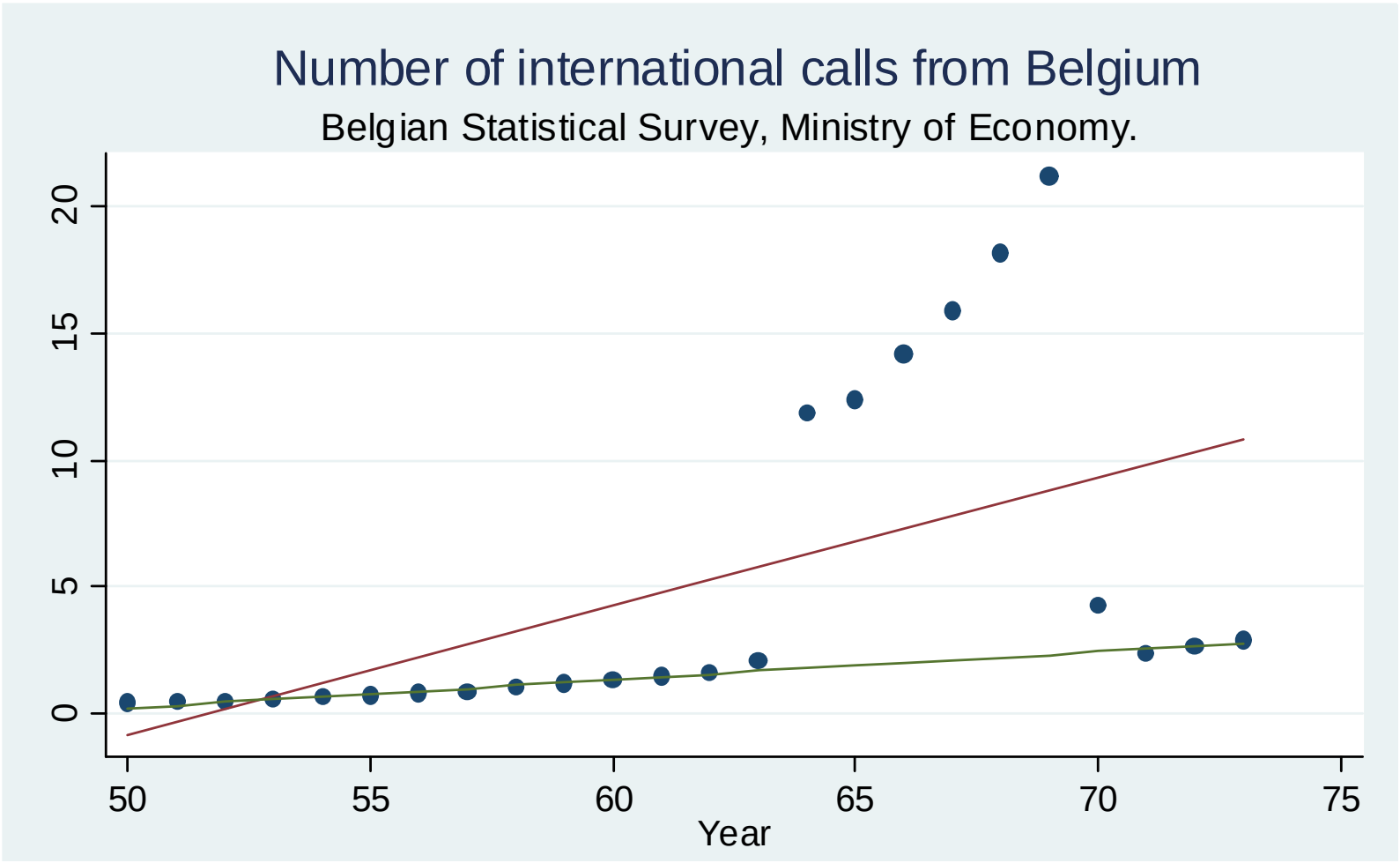
Source: P.J. Rousseeuw and A.M. Leroy (1987)

Outliers do matter and are not necessarily coding errors



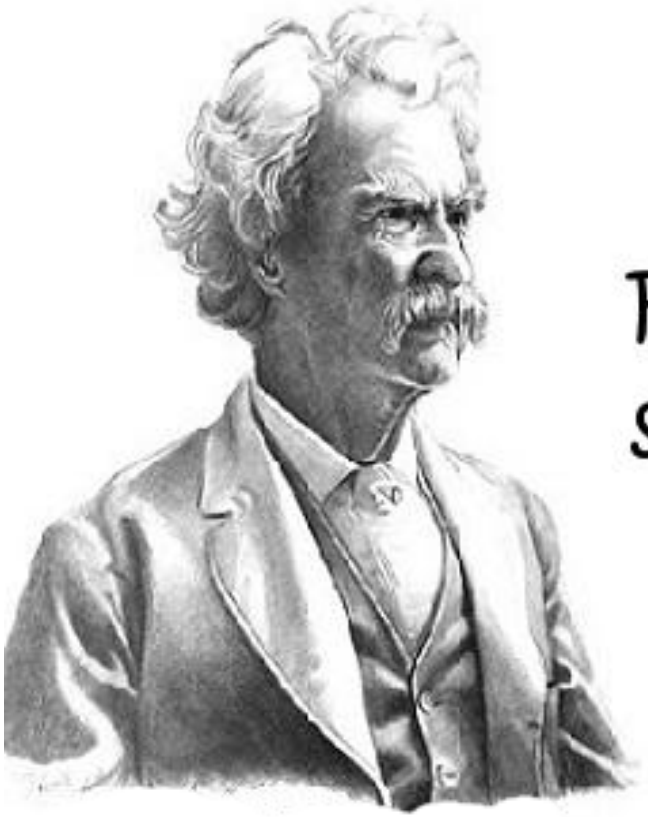
Source: Weisberg, S. (1985)

Outliers do matter and are not necessarily coding errors



Source: P.J. Rousseeuw and A.M. Leroy (1987)

Measuring robustness of an estimator



Facts are stubborn, but statistics are more pliable.

Mark Twain

Measuring robustness of an estimator

Sensitivity curve, see inter alia [Maronna et al., 2006]

Let us consider a data set $\mathbf{X}_n = \{x_1, \ldots, x_n\}$ and the statistic $T_n = T_n(x_1, \ldots, x_n)$.

To study the impact of a potential outlier on this statistic, we may analyze the **modification of value** observed for the statistic when we **add an extra data point** x and allow it to **move on the whole line** (from $-\infty$ to $+\infty$).

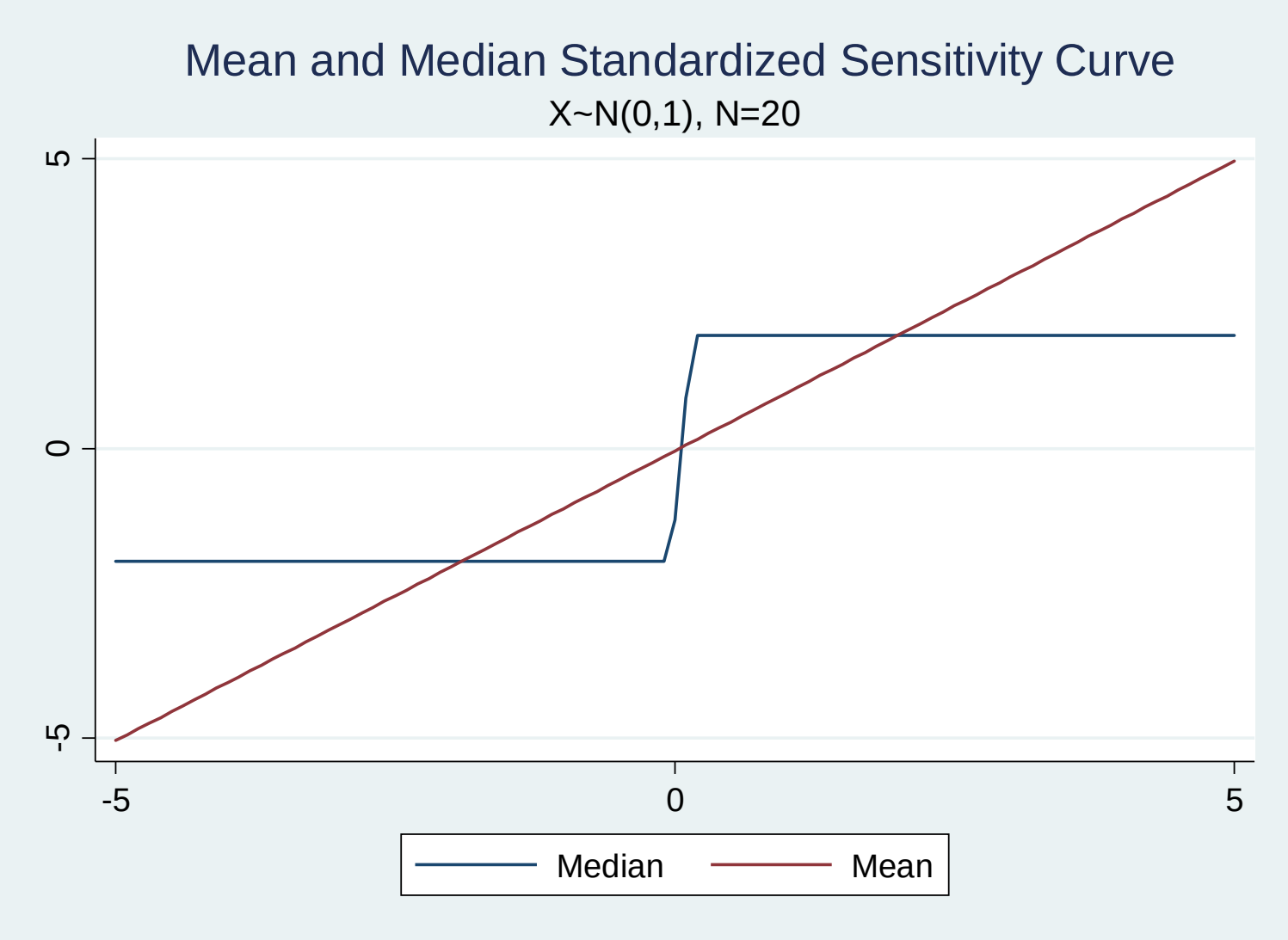
The (standardized) sensitivity curve of the statistic T_n for the sample $\mathbf{X}_n$ is defined by

$$\mathrm{SC}(x; T_n, \mathbf{X}_n) = \frac{T_{n+1}(x_1, \ldots, x_n, x) - T_n(x_1, \ldots, x_n)}{\frac{1}{n+1}};$$

for each value of x, we compare the value of the statistic in the "contaminated" sample with its value in the initial sample, and rescale the difference by dividing by $1/(n+1)$, the amount of contamination.

Measuring robustness of an estimator

Sensitivity curve



Measuring robustness of an estimator

Influence function

The influence function (IF) can be considered as an **asymptotic version of the sensitivity curve** of the statistic T_n when the sample size n grows, that is, when the empirical distribution function F_n tends to the underlying population distribution function F:

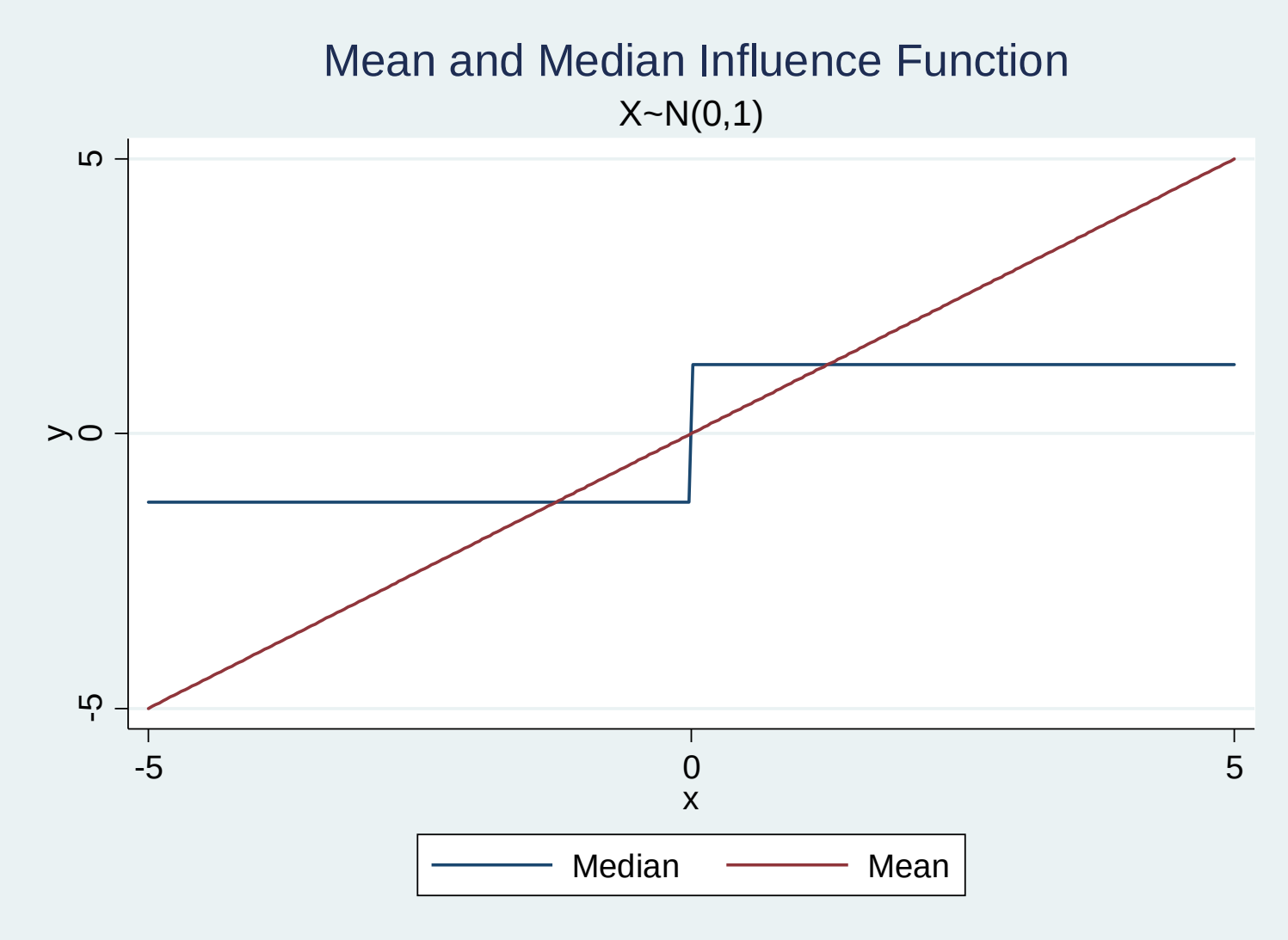
$$\begin{aligned} \mathrm{IF}(x; T, F) &= \lim_{n \to \infty} \frac{T\left(\left(1 - \frac{1}{n+1}\right) F + \frac{1}{n+1} \Delta_x\right) - T(F)}{\frac{1}{n+1}} \\ &= \lim_{\varepsilon \to 0} \frac{T\left((1-\varepsilon) F + \varepsilon \Delta_x\right) - T(F)}{\varepsilon}, \end{aligned}$$

where Δ_x denotes the probability distribution putting all its mass in the point x.

This function measures the effect on T of a pertubation of F obtained by adding a small probability mass at the point x.

Measuring robustness of an estimator

Influence Function



Measuring robustness of an estimator

Gross-error sensitivity

The gross-error sensitivity of T at distribution F, defined by

$$\gamma^*(T, F) = \sup_x |\mathrm{IF}(x; T, F)|,$$

evaluates the **biggest influence that an outlier may have** on T. From the robustness point of view, it is of course preferable to use an estimator for which $\gamma^*(T, F)$ is finite (i.e. bounded IF).

Local-shift sensitivity

The local-shift sensitivity measures the **effect of a small perturbation** of the value of x on T. We may determine the local-shift sensitivity

$$\lambda^*(T, F) = \sup_{x \neq y} \frac{|\mathrm{IF}(y; T, F) - \mathrm{IF}(x; T, F)|}{|y - x|}.$$

From the robustness point of view, it is of course preferable to use an estimator for which the IF is smooth everywhere.

Measuring robustness of an estimator

Breakdown point

The sensitivity curve shows how an estimator reacts to the introduction of one single outlier. Some estimators have bounded sensitivity curve (SC) and therefore resist to this contamination. However, it is possible that the number of outliers in a sample is so large that even these estimators with bounded SC can break.

The **breakdown point** is, roughly, the **smallest amount of contamination** in the sample that may **cause the estimator to take on arbitrary values**.

Example

If the ith observation among $x_1, \ldots, x_n$ goes to infinity, the sample mean μ_n goes to infinity as well. This means that the finite-sample breakdown point of this statistic is only $1/n$. In contrast, the finite-sample breakdown point of the median $Q_{0.5;n}$ is $\frac{n/2}{n}$ if n is even and $\frac{(n+1)/2}{n}$ if n is odd.

Measuring robustness of an estimator

Choosing a good (robust) estimator

- **Fisher-consistent**. If the estimator was calculated using the entire population rather than a sample, the true value of the estimated parameter should be obtained
- **Bounded influence function** (low gross-error sensitivity). The biggest influence that an outlier may have on the estimator should be limited
- **Smooth influence function** (low local-shift sensitivity). The effect on the estimator of a small perturbation in the data should be limited
- **High breakdown point**. The estimator must withstand a contamination of a large proportion of the data
- **Highly efficient** with *convergence rate of* $\sqrt{n}$
- **Computationally feasible**

Compromises must often be made to achieve good performance.

Descriptive statistics

Without data you're just another person with an opinion

W. Edwards Deming

Descriptive statistics

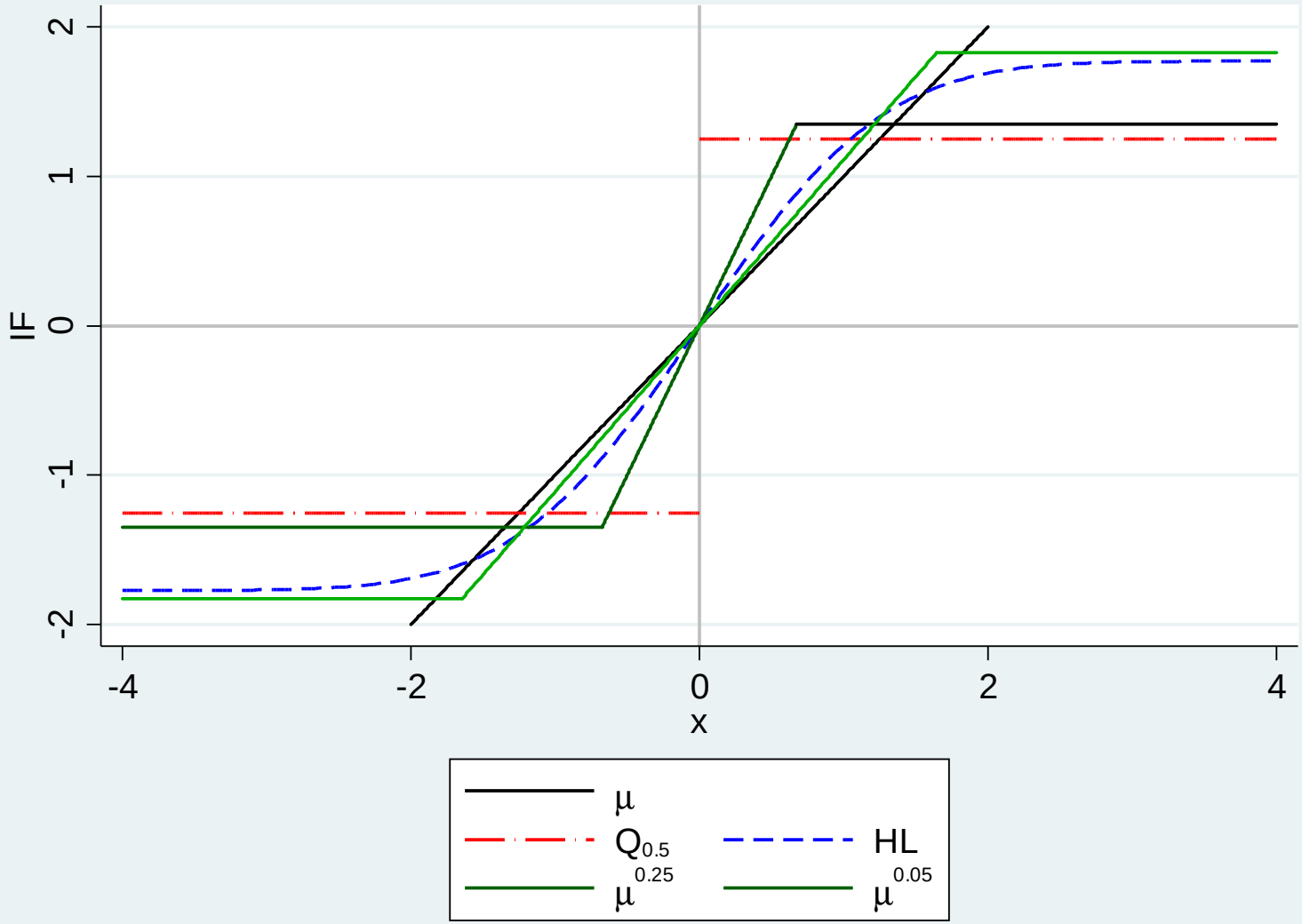
Location parameters

Several measures of location are available in the literature. We compare i) two "classical" estimators based on (centered) moments of the empirical distribution, ii) an estimator based on quantiles of the distribution, and iii) an estimator based on pairwise comparisons of the observations

- **Classical estimator (mean)**
  $\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Classical estimator (trimmed mean)**
  $\mu_n^{\alpha} = \frac{1}{n-2\lfloor \alpha n \rfloor}\sum_{i=\lfloor \alpha n \rfloor+1}^{n-\lfloor \alpha n \rfloor} x_{(i)}$
- **Quantile-based estimator (median)**
  $\mathrm{Q}_{0.5} = \mathrm{med}\left\{x_i\right\}$
- **Pairwise based estimator** [Hodges and Lehmann, 1963]
  $\mathrm{HL}_n = \mathrm{med}\left\{\frac{x_i+x_j}{2}; i<j\right\}$

Location parameters

Influence functions



Location parameters

Comparing properties of location estimators

Estimator	$\mathrm{ASV}(\cdot, \Phi)$	Asymptotic breakdown value	Computational complexity
μ_n	1	0%	$O(n)$
μ_n^α	$\begin{cases} 1.0263 & \text{if } \alpha = 0.05 \\ 1.0604 & \text{if } \alpha = 0.10 \\ 1.1952 & \text{if } \alpha = 0.25 \end{cases}$	$100\alpha\%$	$O(n)$
$Q_{0.5;n}$	$\pi/2 = 1.5708$	50%	$O(n)$
HL_n	$\pi/3 = 1.0472$	29%	$O(n \log n)$

Stata example

Clean dataset

```
clear
set seed 1234
set obs 10000
drawnorm z
gen x=z
sum x, d
robstat x, stat(hl)
```

	μ_n	$Q_{0.5;n}$	HL_n
Value	-0.00	-0.01	-0.01
Time	0.01	0.01	0.40

Contaminated dataset

```
clear
set seed 1234
set obs 10000
drawnorm z
gen x=z+10 in 1/100
sum x, d
robstat x, stat(hl)
```

	μ_n	$Q_{0.5;n}$	HL_n
Value	1.00	0.00	0.01
Time	0.01	0.01	0.41

Descriptive statistics

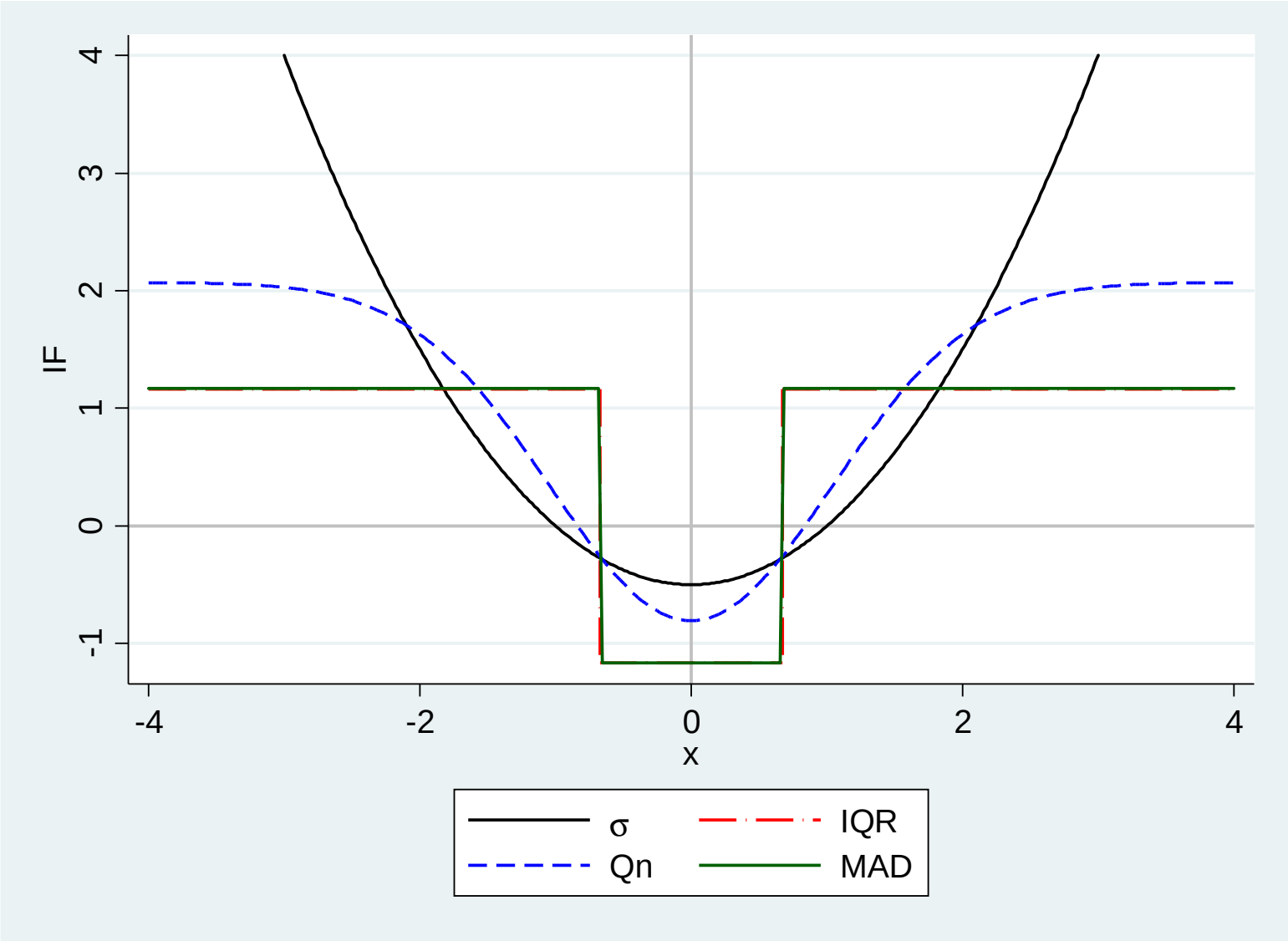
Scale parameters

Several measures of dispersion are available in the literature. We compare i) a "classical" estimators based on (centered) moments of the empirical distribution, ii) two estimators based on quantiles of the distribution, and iii) an estimator based on pairwise comparisons of the observations

- **Classical estimator (standard deviation)**
  $\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_n)^2}$
- **Quantile-based estimator (inter-quartile range)**
  $\mathrm{IQR}_n = 0.7413 \times (Q_{0.75} - Q_{0.25})$
- **Quantile-based estimator (median absolute deviation)**
  $\mathrm{MAD}_n = 1.4826 \times \mathrm{med}_i |x_i - \mathrm{med}_j x_j|$
- **Pairwise based estimator [Rousseeuw and Croux, 1993]**
  $Q_n = 2.2219 \times \{|x_i - x_j|; i < j\}_{(k)}$ and $k = \binom{n}{2}/4$

Scale parameters

Influence functions



Scale parameters

Comparing properties of location estimators

Estimator	Type	$\mathrm{ASV}(\cdot, \Phi)$	Asymptotic breakdown value	Computational Complexity
σ_n	Classical	0.5	0%	$O(n)$
IQR_n	Quantile-based	1.3605	25%	$O(n)$
MAD_n	Quantile-based	1.3605	50%	$O(n)$
Q_n	Pairwise-based	0.6077	50%	$O(n \log n)$

Scale parameters

Stata example

Clean dataset

```
clear
set seed 1234
set obs 10000
drawnorm z
gen x=z
robstat, stat(sd iqrc)
robstat, stat(qn)
```

	σ_n	IQR_n	Q_n
Value	0.99	1.00	1.00
Time	0.02	0.07	0.58

Contaminated dataset

```
clear
set seed 1234
set obs 10000
drawnorm z
gen x=z+10 in 1/100
robstat, stat(sd iqrc)
robstat, stat(qn)
```

	σ_n	IQR_n	Q_n
Value	10.00	1.00	1.02
Time	0.01	0.10	0.52

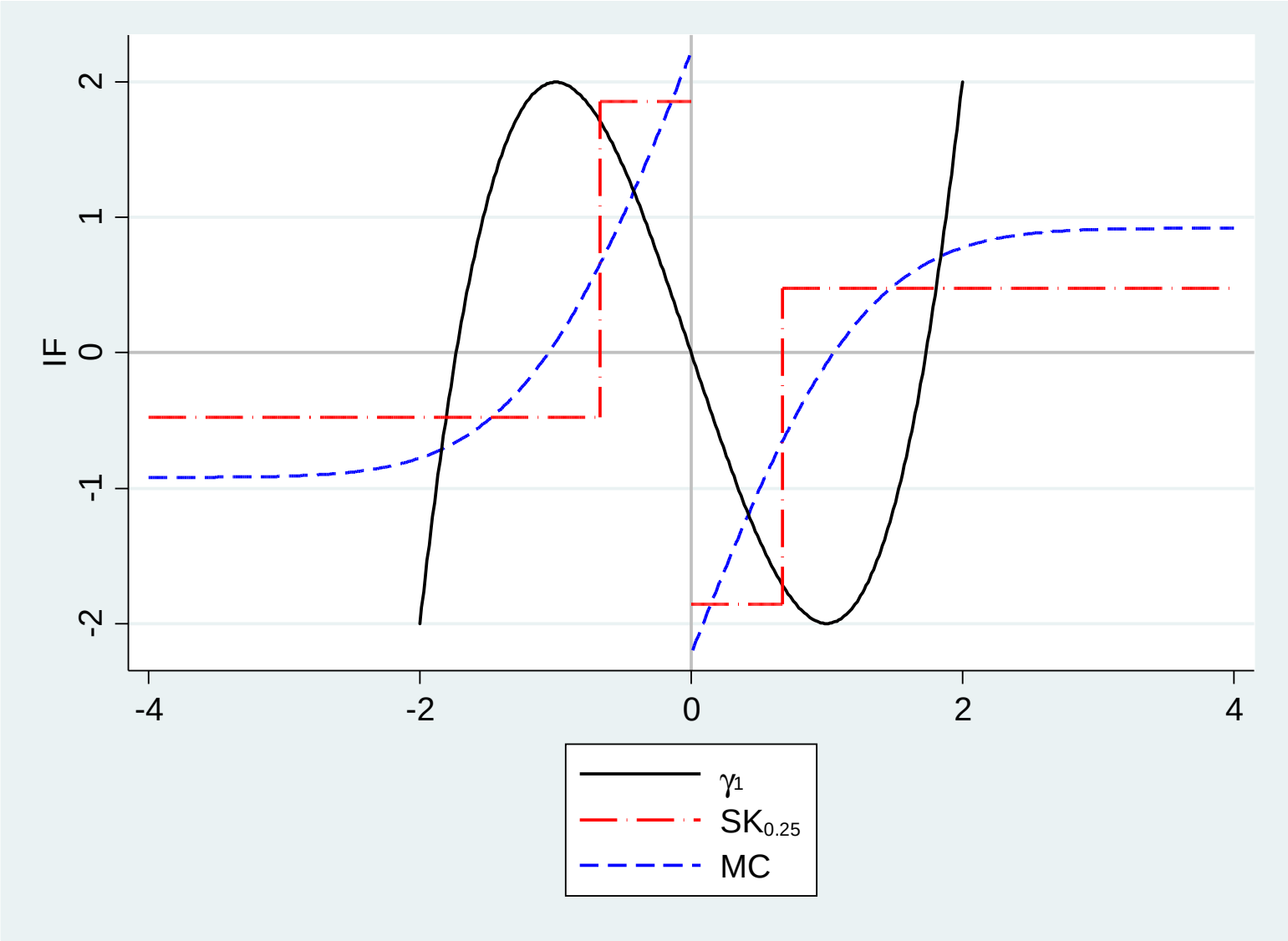
Skewness parameters

Several measures of skewness are available in the literature. We compare i) a "classical" estimators based on (centered) moments of the empirical distribution, ii) an estimators based on quantiles of the distribution, and iii) an estimator based on pairwise comparisons of the observations

- **Classical estimator (Fisher coefficient)**
  $\gamma_{1;n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu_n}{\sigma_n}\right)^3$
- **Quantile-based estimator (Hinkley $p = 0.25$)**
  $\mathrm{SK}_{0.25;n} = \frac{(Q_{1-p}-Q_{0.5})-(Q_{0.5}-Q_p)}{Q_{1-p}-Q_p} = \frac{Q_p+Q_{1-p}-2Q_{0.5}}{Q_{1-p}-Q_p}$
- **Pairwise-based estimator-Medcouple [Brys et al., 2004]**
  $\mathrm{MC}_n = \mathrm{med}_{x_{(i)}\leq Q_{0.5;n}\leq x_{(j)}} \frac{\left(x_{(j)}-Q_{0.5;n}\right)-\left(Q_{0.5;n}-x_{(i)}\right)}{x_{(j)}-x_{(i)}}$ for all $x_{(i)} \neq x_{(j)}$

Skewness parameters

Influence functions



Comparing properties of skewness estimators

Estimator	Type	$\mathrm{ASV}(\cdot, \Phi)$	Asymptotic breakdown value	Computational complexity
$\gamma_{1;n}$	Classical	6	0%	$O(n)$
$\mathrm{SK}_{0.25;n}$	Quantile-based	1.8421	25%	$O(n)$
MC_n	Pairwise-based	1.25	25%	$O(n \log n)$

Skewness parameters

Stata example

Clean dataset

```
clear
set seed 1234
set obs 10000
drawnorm z
gen x=z
robstat x, stat(skew sk)
robstat x, stat(mc)
```

	$\gamma_{1;n}$	$SK_{0.25;n}$	MC_n
Value	0.02	0.00	0.00
Time	0.00	0.09	0.56

Contaminated dataset

```
clear
set seed 1234
set obs 10000
drawnorm z
gen x=z+10 in 1/100
robstat x, stat(skew sk)
robstat x, stat(mc)
```

	$\gamma_{1;n}$	$SK_{0.25;n}$	MC_n
Value	9.70	0.01	0.01
Time	0.00	0.12	0.56

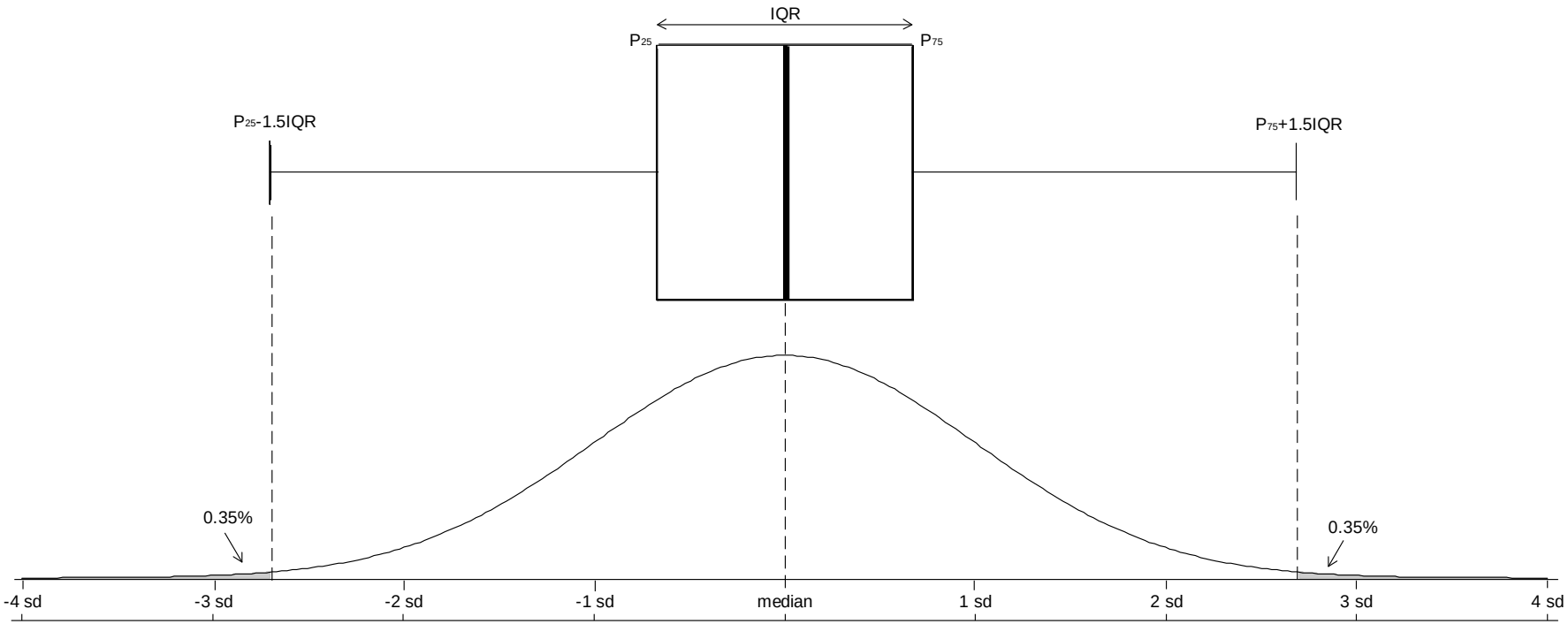
Identifying univariate outliers

The truth is rarely pure and never simple.

Oscar Wilde

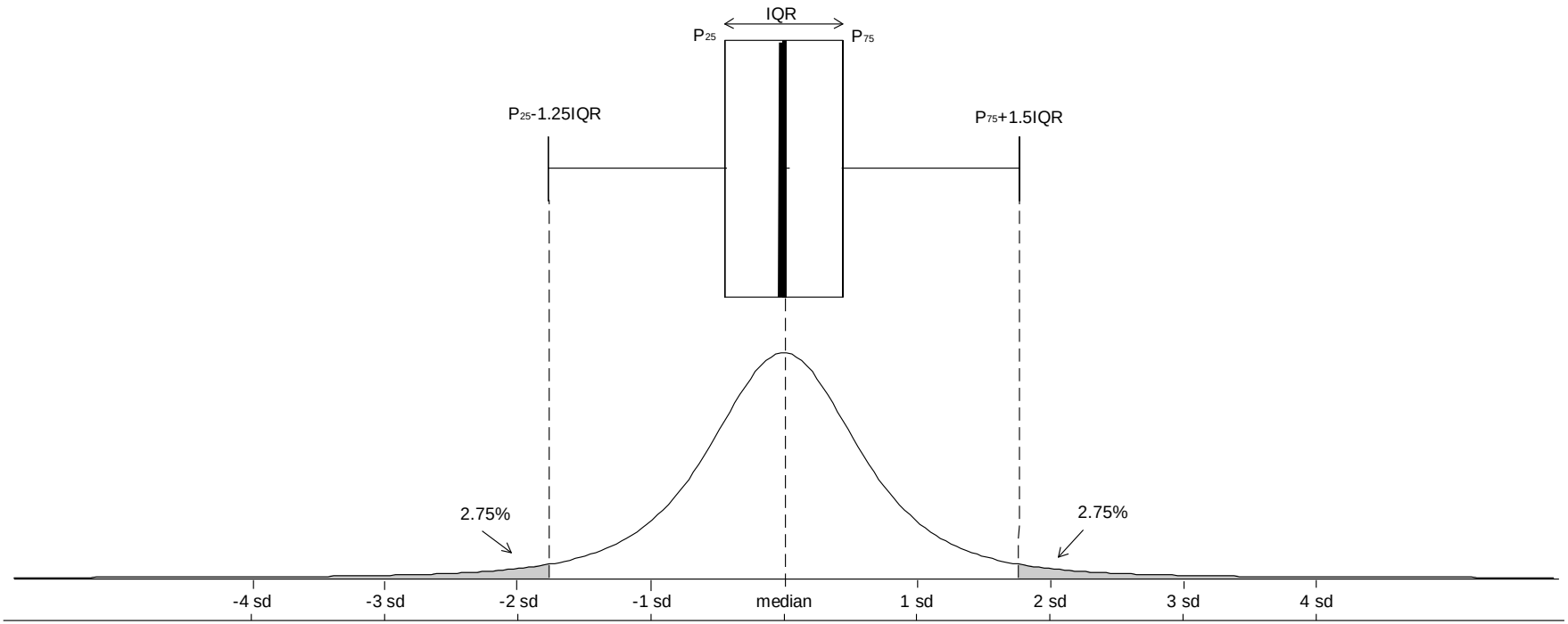
Univariate outliers identification

Standard Boxplot, Standard Normal distribution



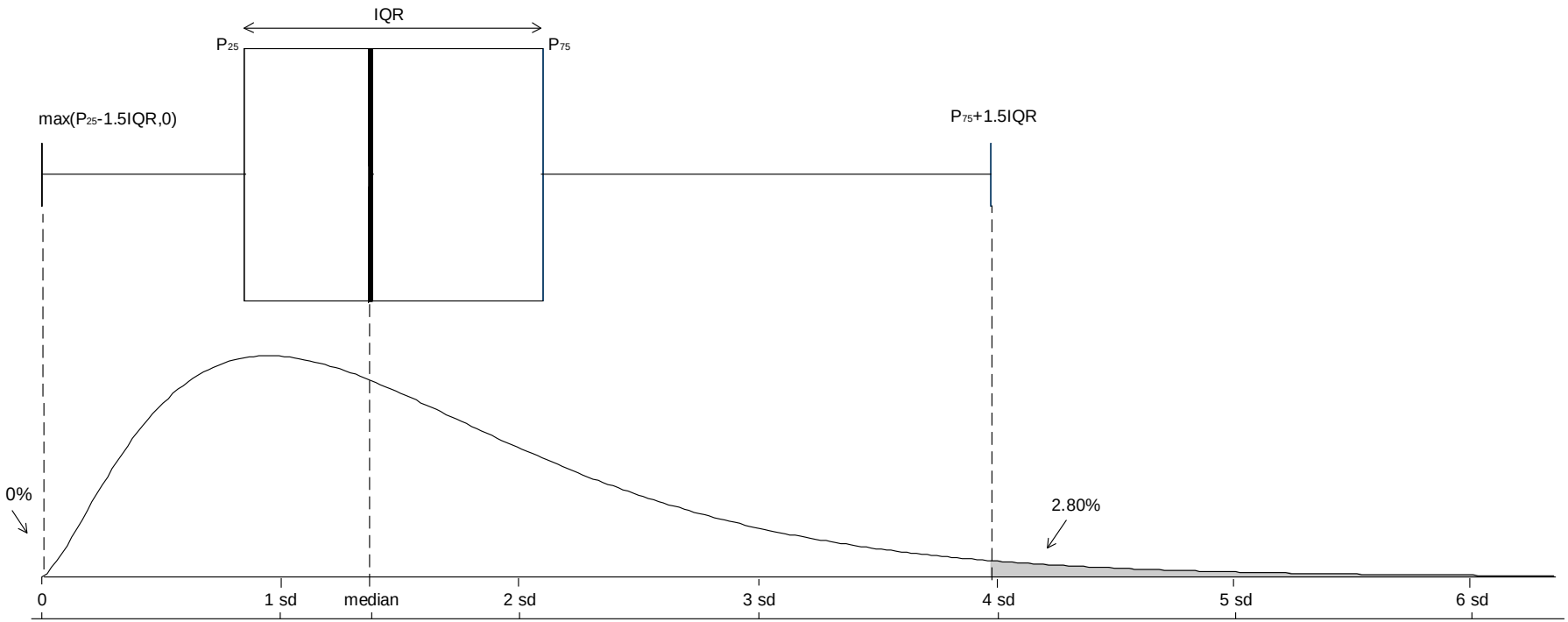
Univariate outliers identification

Standard Boxplot, heavy tailed t_2 distribution



Univariate outliers identification

Standard Boxplot, skewed χ^2_5 distribution



Univariate outliers identification

Must adjust to skewness and tail heavyness [Bruffaerts et al., 2014]

Modify the whiskers of the boxplot to deal with asymmetry and tail heavyness.

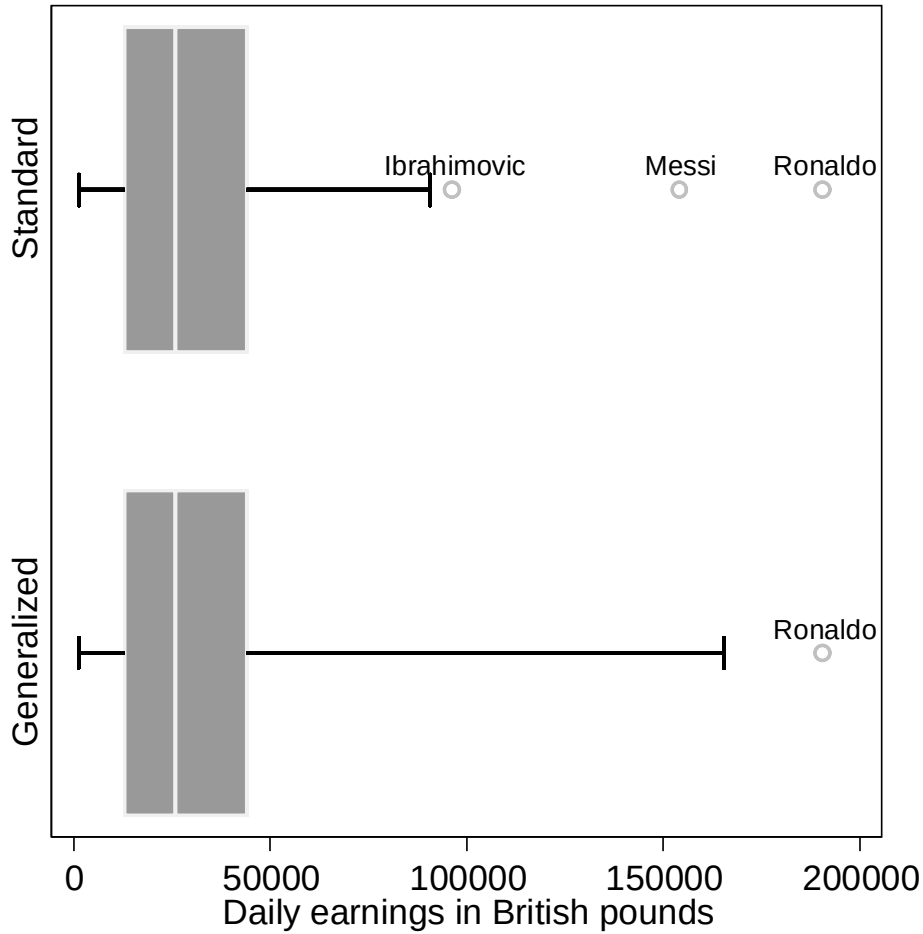
Generalized Boxplot

- **Cope with** both the **skewness** and **tail heavyness**
- **Set** the desired **rejection rate** to any chosen level
- **Computational complexity** $O(n)$
- Do a **rank preserving transformation** of the data
- **Fit** the transformed data density using a **Tukey g and h** distribution
- Chose the theoretical quantiles of the latter to **set whiskers** (after applying an inverse transformation)

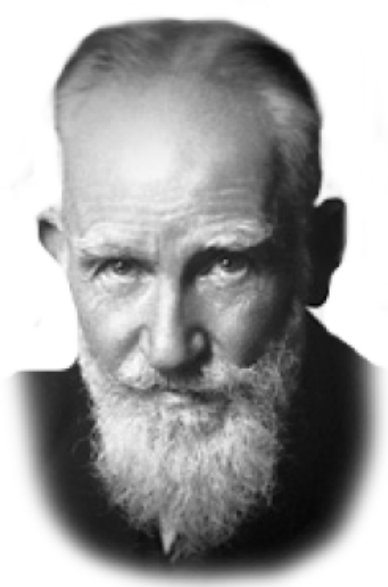
Univariate outliers identification

Must adjust to skewness and tail heavyness

- `box_out x, gen(out)`



Regression models

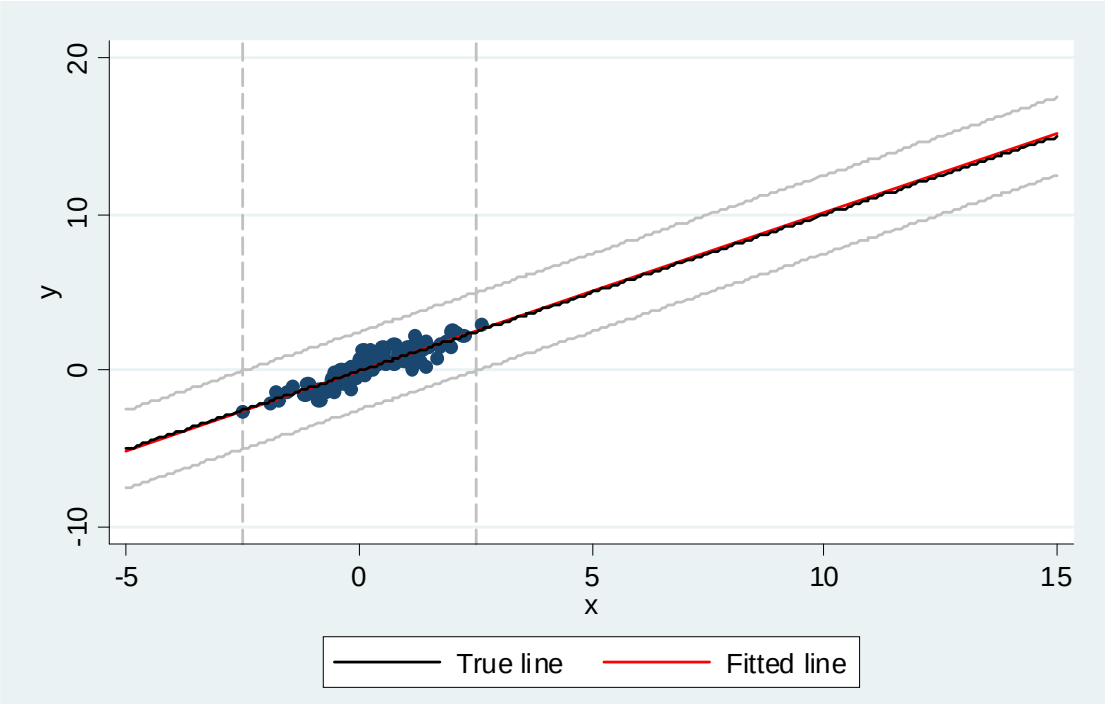


Science never solves a problem without creating ten more

George Bernard Shaw

Classical modelling: outliers typology

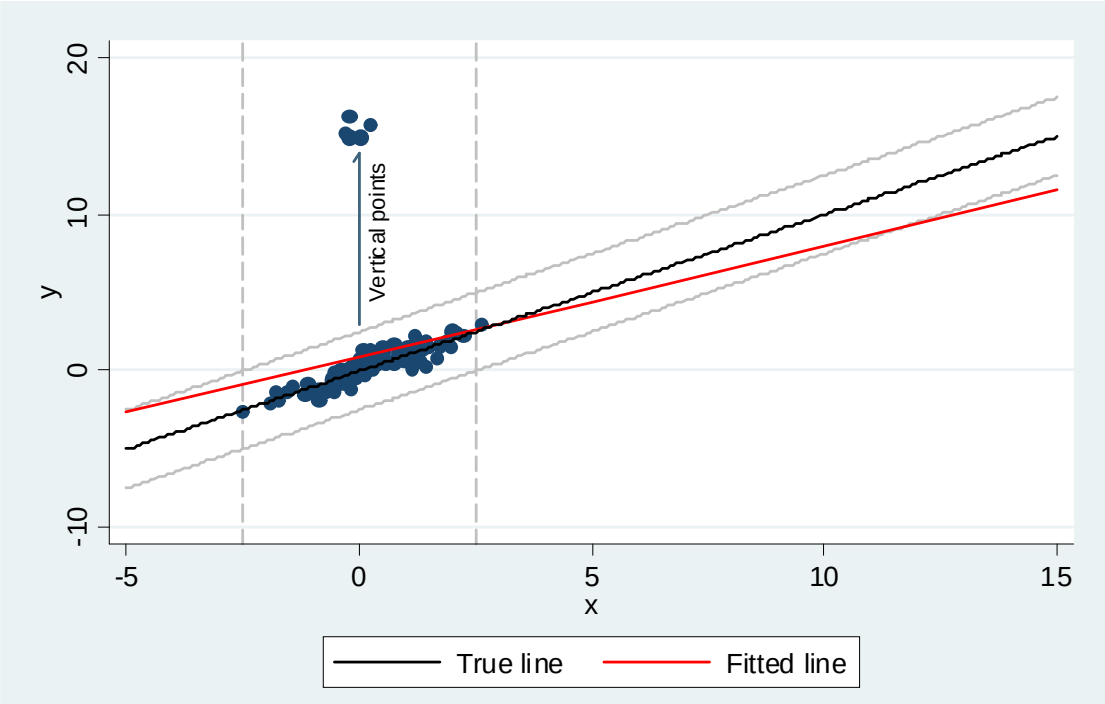
Influential outliers



	Clean	Vertical	Bad leverage	Good leverage
Intercept	0.012			
t-stat	(0.22)			
Slope	1.013			
t–stat	(19.49)			

Classical modelling: outliers typology

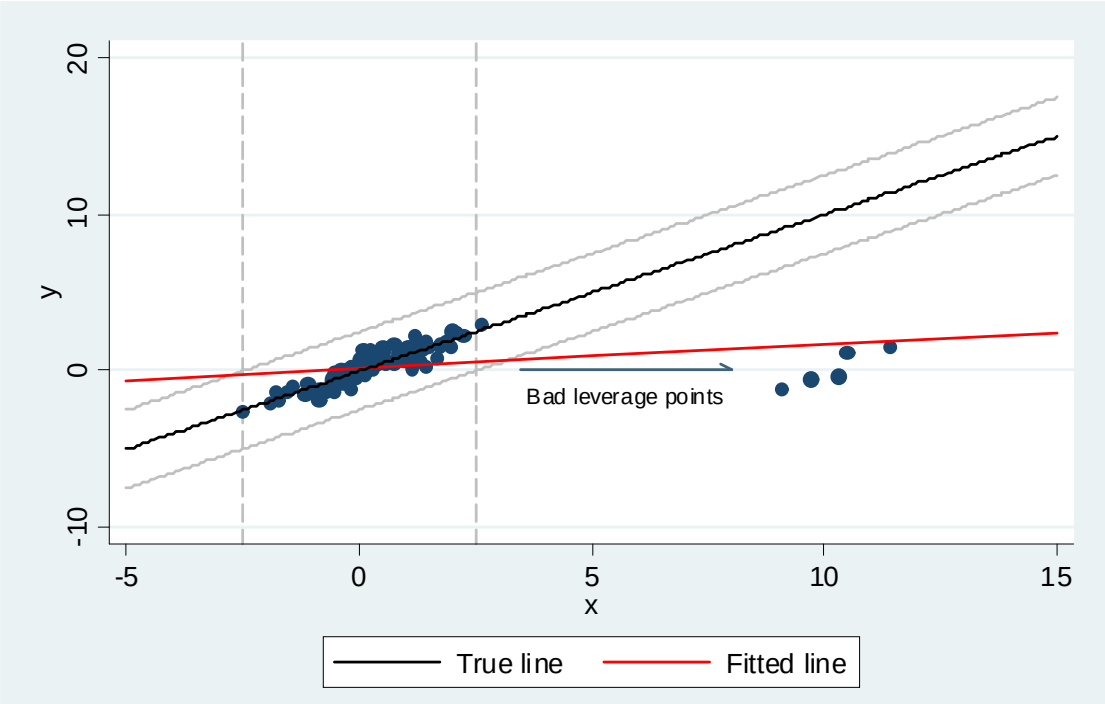
Influential outliers



	Clean	Vertical	Bad leverage	Good leverage
Intercept	0.012	0.964		
t-stat	(0.22)	(2.44)		
Slope	1.013	0.704		
t–stat	(19.49)	(1.88)		

Classical modelling: outliers typology

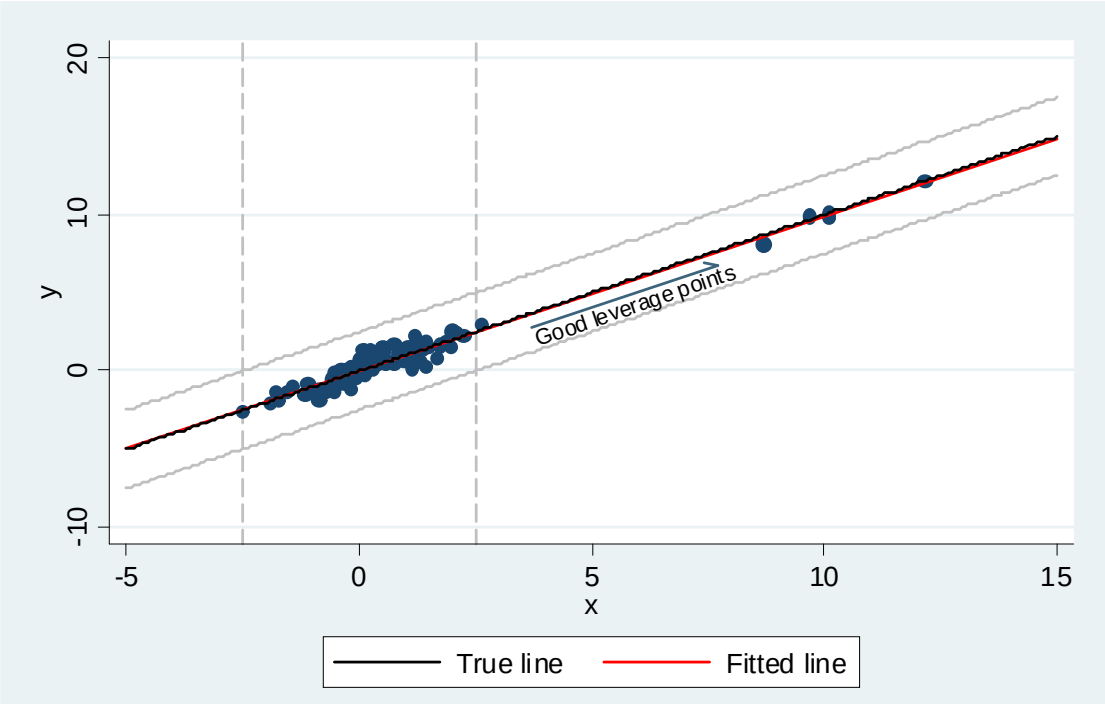
Influential outliers



	Clean	Vertical	Bad leverage	Good leverage
Intercept	0.012	0.964	0.163	
t-stat	(0.22)	(2.44)	(1.33)	
Slope	1.013	0.704	0.155	
t–stat	(19.49)	(1.88)	(3.34)	

Classical modelling: outliers typology

Influential outliers



	Clean	Vertical	Bad leverage	Good leverage
Intercept	0.012	0.964	0.163	0.016
t-stat	(0.22)	(2.44)	(1.33)	(0.29)
Slope	1.013	0.704	0.155	0.988
t–stat	(19.49)	(1.88)	(3.34)	(47.92)

Classical modelling: Least Squares regression and L1

Least squares regression

Consider regression model

$$y_i = x_i^t \theta + \varepsilon_i$$

where y_i is the dependent variable, x_i is the vector of covariates and ε_i is the error term $(i = 1, ..., n)$.

To estimate θ, a loss function of the residuals $r_i(\theta) = y_i - x_i^t \theta$ is minimized.

- **LS-estimator:** $\hat{\theta}_{LS} = \arg\min_{\theta} \sum_{i=1}^{n} r_i(\theta)^2$ (`regress` in Stata)

The squaring of the residuals makes LS very sensitive to outliers. To increase robustness, the square function could be replaced by the absolute value [Edgeworth, 1887].

- **L_1-estimator:** $\hat{\theta}_{LS} = \arg\min_{\theta} \sum_{i=1}^{n} |r_i(\theta)|$ (`qreg` in Stata)

Robust modelling: M-estimators

M-estimators

[Huber, 1981] generalized this idea to a set of symmetric functions that could be used instead of the absolute value to increase efficiency and robustness.

To guarantee scale equivariance, residuals are standardized by a measure of dispersion σ.The problem becomes:

- **M-estimator:** $\hat{\theta}_M = \arg\min_{\theta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\theta)}{\sigma}\right)$ (`robreg m` in Stata)

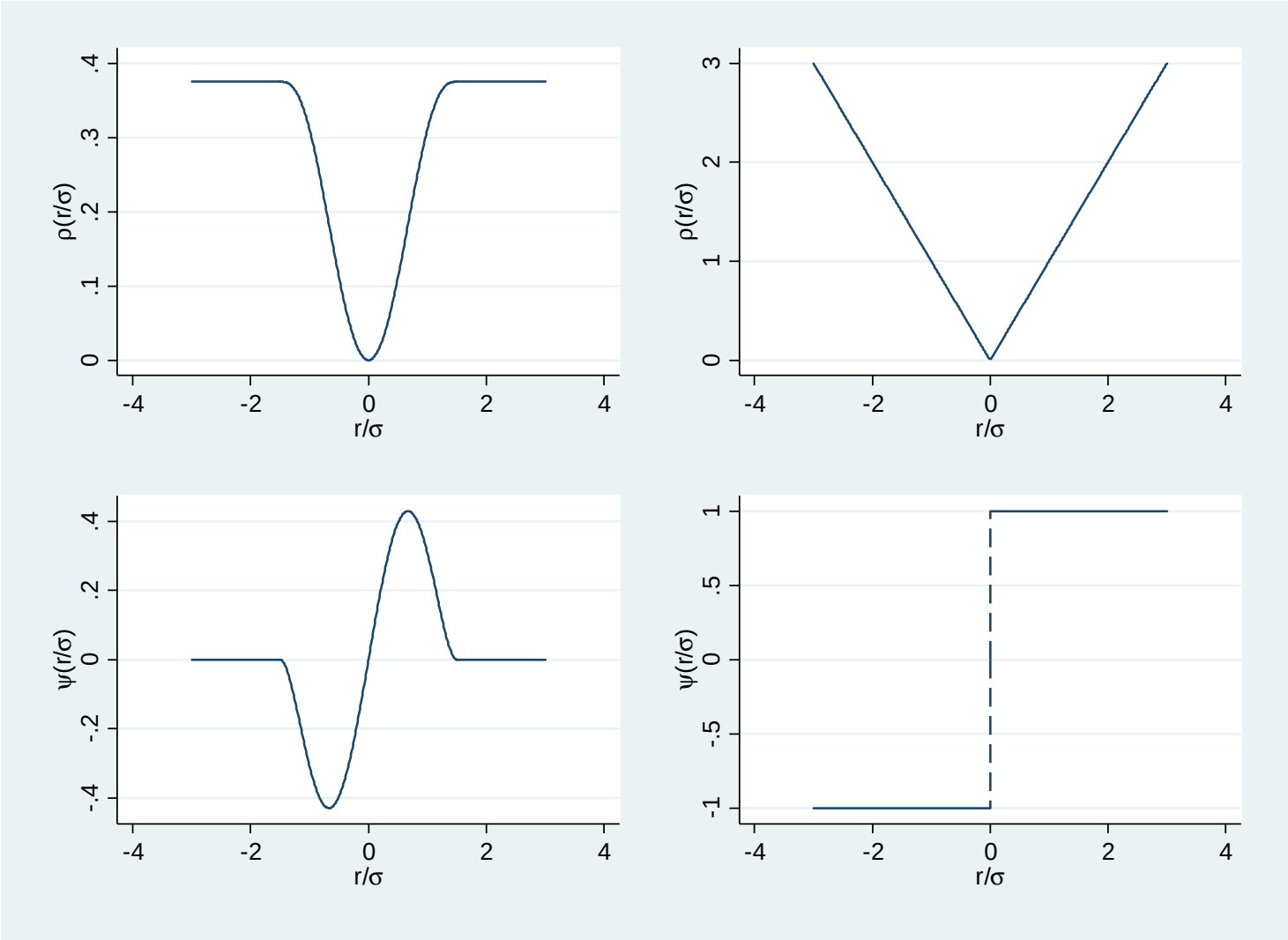
The function ρ or its derivative, ψ, can be chosen in such a way to provide the estimator desirable properties in terms of bias and efficiency.

For many choices of ρ or ψ, no closed form solution exists. In most cases an iteratively re-weighted least squares fitting algorithm can be performed.

If the function ψ decreases to zero as $\frac{r_i(\theta)}{\sigma} \longrightarrow \pm\infty$, the estimator is called redescending.

Robust modelling: M-estimators

M-estimators can be redescending (left) or monotonic (right).



Robust modelling: pitfalls of M-estimators

M-estimator: $\hat{\theta}_M = \arg\min_{\theta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\theta)}{\sigma}\right)$

- If σ **is known beforehand**, setting $w_i = \rho\left(\frac{r_i(\theta)}{\sigma}\right) \big/ r_i^2(\theta)$ we have that $\hat{\theta}_M = \arg\min_{\theta} \sum_{i=1}^{n} w_i r_i^2(\theta)$ that can be easily fitted using iteratively reweighted least squares.
- The **convergence** of the algorithm to a unique solution is **guaranteed for monotonic estimators** but nor for redescending M-estimators.
- σ is not known beforehand and has to be estimated which is not easy without a **good starting point**.
- **Monotonic M-estimators are not robust** against bad leverage points. Indeed F.O.C $-\sum_{i=1}^{n} \underbrace{\psi\left(\frac{r_i(\theta)}{\sigma}\right)}_{\to 0} \underbrace{x_i}_{\to \infty} = 0$.
- The problem must be tackled from a **different perspective**

S-estimator of regression

The square function in LS awards excessive importance to outliers. To increase robustness, another function $\rho_0(\cdot)$ (even, non decreasing for positive values, less increasing than the square with a minimum at zero) should be preferred

- **LS-estimator:** $\begin{cases} \min\limits_{\theta} s(r_1(\theta), ..., r_n(\theta)) \\ \text{s.t.} \frac{1}{n} \sum\limits_{i=1}^{n} \left(\frac{y_i - x_i^t \theta}{s} \right)^2 = 1 \end{cases}$
- **S-estimator:** $\begin{cases} \min\limits_{\theta} s(r_1(\theta), ..., r_n(\theta)) \\ \text{s.t.} \frac{1}{n} \sum\limits_{i=1}^{n} \rho_0 \left(\frac{y_i - x_i^t \theta}{s} \right) = \delta \end{cases}$
- where $\delta = E\left[\rho_0(u)\right]$ with $u \backsim N(0,1)$

S-estimator of regression

The optimization problem can therefore be written as

- **S-estimator:** $\begin{cases} \min_{\theta} s(r_1(\theta), ..., r_n(\theta)) \\ \text{s.t.} \frac{1}{n} \sum_{i=1}^{n} \rho_0 \left(\frac{y_i - x_i^t \theta}{s} \right) = \delta \end{cases}$

and the associated first order conditions are

- **F.O.C**: $\begin{cases} \frac{1}{n} \sum_{i=1}^{n} \rho_0' \left(\frac{y_i - x_i^t \hat{\theta}_0}{\hat{\sigma}} \right) x_i^t = 0 \\ \frac{1}{n} \sum_{i=1}^{n} \rho_0 \left(\frac{y_i - x_i^t \hat{\theta}_0}{\hat{\sigma}} \right) = \delta \end{cases}$

where ρ_0' is the first derivative of ρ_0

Robust modelling: S-estimator of regression

Tukey Biweight Function

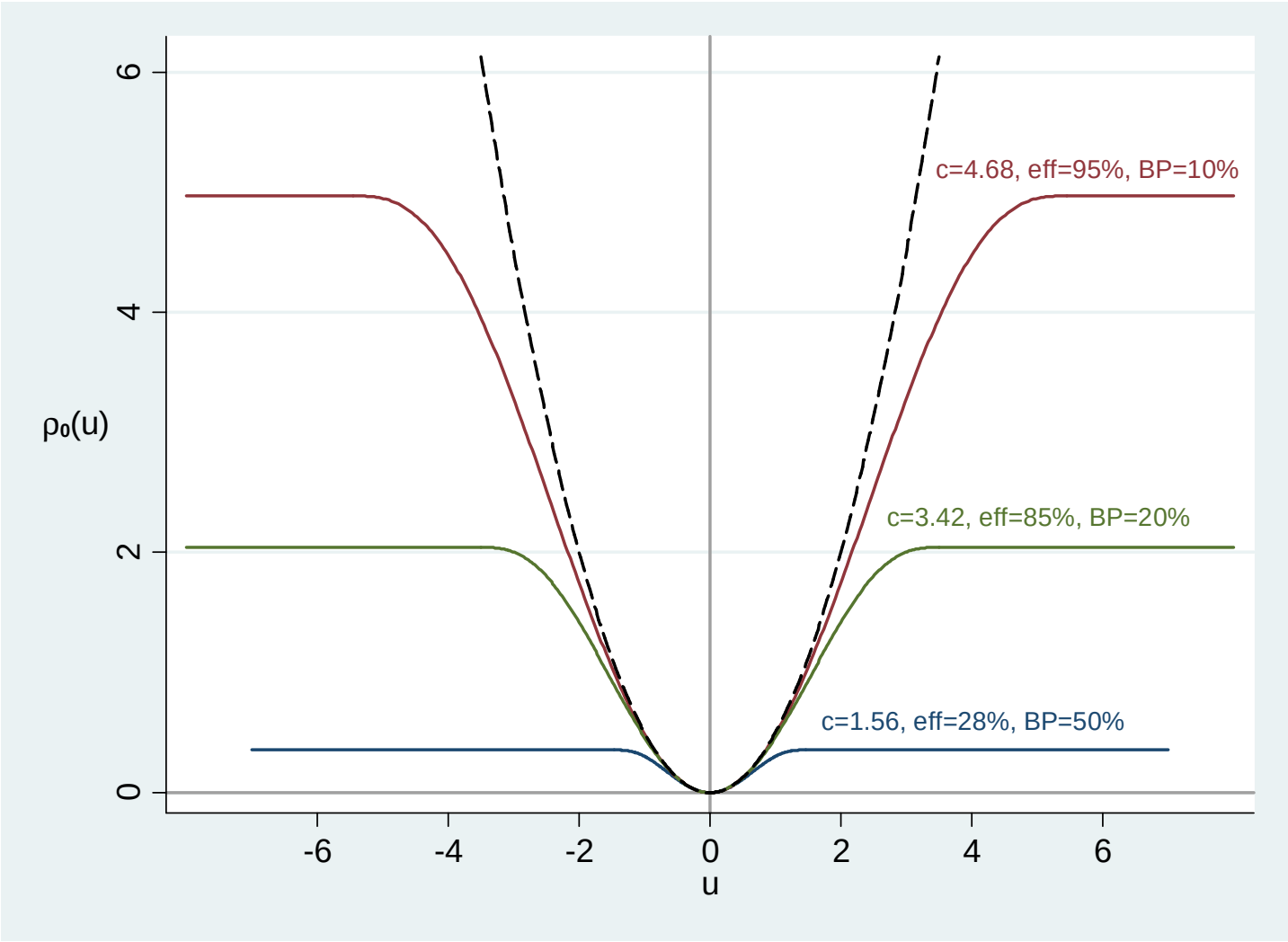
Several ρ_0 functions can be used. We chose Tukey's Biweight function here defined as

$$\rho_0(u) = \begin{cases} \frac{c^2}{6}\left(1 - \left[1 - \left(\frac{u}{c}\right)^2\right]^3\right) & \text{if } |u| \leq c \\ \frac{c^2}{6} & \text{if } |u| > c \end{cases} . \tag{1}$$

There is a **trade-off** between **robustness** and Gaussian **efficiency**

- $c = 1.56$ leads to a 50% BD and an efficiency of 28%
- $c = 3.42$ leads to a 20% BDP and an efficiency of 85%
- $c = 4.68$ leads to a 10% BDP and an efficiency of 95%

Tukey Biweight Function



Robust estimators

MM-estimators [Yohai, 1987]

1. **Fit an S-estimator** of regression with **50% BDP** and estimate the scale parameter $\hat{\sigma}_S = s(r_1(\hat{\theta}_S), \ldots, r_n(\hat{\theta}_S))$.
2. Take another function $\rho \geq \rho_0$ and **estimate:** $\hat{\theta}_{MM} = \arg\min_{\theta} \sum_{i=1}^{n} \rho(\frac{r_i(\theta)}{\hat{\sigma}_S})$. The BDP is set by ρ_0 and the efficiency by ρ.
3. The **first order conditions** associated to the MM-estimator are

$$\textbf{F.O.C} \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{y_i - x_i^t\hat{\theta}}{\hat{\sigma}}\right) x_i^t = 0 & \text{MM} \\ \boxed{\begin{array}{l} \frac{1}{n}\sum_{i=1}^{n} \rho_0'\left(\frac{y_i - x_i^t\hat{\theta}_0}{\hat{\sigma}}\right) x_i^t = 0 \\ \frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{y_i - x_i^t\hat{\theta}_0}{\hat{\sigma}}\right) = \delta \end{array}} & \text{S} \end{cases}$$

where ψ is the first derivative of ρ.

GMM estimator

[Croux et al., 2003] suggest that MM-estimate are equivalent to exactly-identified GMM estimators for $\vartheta = (\theta^t, \theta_0^t, \sigma)^t$ with moment matrix

$$m_i(\hat{\vartheta}) = \begin{pmatrix} \psi\left(\frac{y_i - x_i^t\hat{\theta}}{\hat{\sigma}}\right) x_i^t \\ \rho_0'\left(\frac{y_i - x_i^t\hat{\theta}_0}{\hat{\sigma}}\right) x_i^t \\ \rho_0\left(\frac{y_i - x_i^t\hat{\theta}_0}{\hat{\sigma}}\right) - \delta \end{pmatrix}$$

As shown in [Hansen, 1982] $\hat{\vartheta}$ has a limiting normal distribution given by

$$\sqrt{n}(\hat{\vartheta} - \vartheta) \longrightarrow N(0, V)$$

where, defining $G = E\left[\frac{\partial m_i(\vartheta)}{\partial \vartheta^t}\right]$ and $\Omega = E[m_i(\vartheta) m_i^t(\vartheta)]$, the asymptotic variance is $V = G^{-1}\Omega(G^t)^{-1}$

Choosing the estimator

Two Hausman-type tests [Dehon et al., 2012]

Question 1: Generalized Hausman-Type test (GH), LS is a particular MM-estimator when $c_1 \longrightarrow \infty$

$$GH = (\hat{\theta}_S - \hat{\theta}_{LS})^t [Var(\hat{\theta}_S) + Var(\hat{\theta}_{LS}) - 2Cov(\hat{\theta}_{LS}, \hat{\theta}_S)]^{-1} (\hat{\theta}_S - \hat{\theta}_{LS})$$

Under the null, GH is distributed asymptotically as a central χ^2_p where p is the number of unknown parameters. This **test** allows to determine if a **robust method should be preferred** to a classical one.

Question 2: Compare MM and S-estimators

$$GH = (\hat{\theta}_S - \hat{\theta}_{MM})^t [Var(\hat{\theta}_S) + Var(\hat{\theta}_{MM}) - 2Cov(\hat{\theta}_{MM}, \hat{\theta}_S)]^{-1} (\hat{\theta}_S - \hat{\theta}_{MM})$$

Under the null, GH is distributed asymptotically as a central χ^2_p where p is the number of unknown parameters. This **test** allows to figure out the **"optimal" efficiency**

Choosing the estimator

Obesity prevalence in a sample of US counties

```
use
"C:\Users\VV\Dropbox\Vermandele\Multivariate\old\Diabetes.dta
clear
set seed 1234567
gen sort=uniform()
sort sort
keep in 1/500
reg perdiabet percphys perco
robreg s perdiabet percphys perco, hausman
robreg mm perdiabet percphys perco, hausman
robreg mm perdiabet percphys perco, eff(60) hausman
```

Choosing the estimator

Linear regression

```
. reg perdiabet percphys perco
```

Source	SS	df	MS
Model	2171.32424	2	1085.66212
Residual	1311.76184	497	2.63935983
Total	3483.08608	499	6.98013242

Number of obs	= 500
F(2, 497)	= 411.34
Prob > F	= 0.0000
R-squared	= 0.6234
Adj R-squared	= 0.6219
Root MSE	= 1.6246

perdiabet	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
percphys	.1718203	.0190935	9.00	0.000	.1343063	.2093343
percob	.3003667	.0232706	12.91	0.000	.2546458	.3460876
_cons	-3.265407	.507357	-6.44	0.000	-4.262236	-2.268578

Choosing the estimator

S-estimator and outlier test

```
. robreg s perdiabet percphys perco, hausman nodots
enumerating 50 candidates ... done
refining 2 best candidates ... done

S-Regression (28.7% efficiency)                 Number of obs    =       500
                                                Subsamples       =        50
                                                Breakdown point  =        50
                                                Bisquare k       =  1.547645
                                                Scale estimate   =  1.5505239
```

perdiabet	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
percphys	.2316495	.0239441	9.67	0.000	.1847199	.278579
percob	.1825398	.0320588	5.69	0.000	.1197057	.245374
_cons	-1.425621	.7096118	-2.01	0.045	-2.816435	-.0348074

```
Hausman test of S against LS:    chi2(2) = 12.104611      Prob > chi2 = 0.0024
```

MM-estimator and efficiency test

```
MM-Regression (85% efficiency)                  Number of obs   =        500
                                                Subsamples      =         50
                                                Breakdown point =         50
                                                M-estimate: k   =  3.4436898
                                                S-estimate: k   =   1.547645
                                                Scale estimate  =  1.5505239
                                                Robust R2 (w)   =  .68991266
                                                Robust R2 (rho) =  .44064435
```

perdiabet	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
percphys	.1808932	.0228649	7.91	0.000	.1360789	.2257075
percob	.2659837	.0318083	8.36	0.000	.2036406	.3283269
_cons	-2.495173	.5904253	-4.23	0.000	-3.652386	-1.337961

Hausman test of MM against S: chi2(2) = 8.4158528 Prob > chi2 = 0.0149

MM-estimator and efficiency test

```
MM-Regression (60% efficiency)                  Number of obs   =        500
                                                Subsamples      =         50
                                                Breakdown point =         50
                                                M-estimate: k   =  2.3666372
                                                S-estimate: k   =   1.547645
                                                Scale estimate  =  1.5505239
                                                Robust R2 (w)   =  .76053069
                                                Robust R2 (rho) =  .34941524
```

perdiabet	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
percphys	.2026506	.0265639	7.63	0.000	.1505864	.2547149
percob	.2308929	.0350134	6.59	0.000	.1622679	.2995178
_cons	-2.047126	.5926382	-3.45	0.001	-3.208676	-.8855769

Hausman test of MM against S: chi2(2) = 4.5984588 Prob > chi2 = 0.1003

Identifying multivariate outliers

"The idea of 10 dimensions might sound exciting, but they would cause real problems if you forget where you parked your car."

Stephen Hawking

Outliers in higher dimensions

Multivariate analysis

Multivariate analysis deals with situations in which several variables are measured on each experimental unit.

The multivariate analysis may pursue different objectives:

- **reduction of dimensionality**: principal components analysis, factor analysis, canonical correlation, ...;
- **estimation of explanatory models**: multivariate linear model, generalized linear models, ...;
- **identification, classification of individuals** (units) on the basis of the values of the various variables: discriminant analysis, automatic classification, outliers identification relying on Mahalanobis-type distances $d_i = \sqrt{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})^t \, \widehat{\boldsymbol{\Sigma}}^{-1} \, (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})}$...

Most of these techniques rely on the prior estimation of some parameters of the underlying multivariate distribution.

Outliers in higher dimensions

Multivariate analysis

Let $\mathcal{X}^{(n)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a random sample of n i.i.d. p-variate observations such that, for all $i = 1, \ldots, n$,

$$\boldsymbol{\mu} = \mathrm{E}(\mathbf{x}_i) \quad \text{and} \quad \boldsymbol{\Sigma} = \mathrm{E}\left[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t\right].$$

The classical estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be obtained by the method of moments (which coincide with the maximum likelihood estimators when the distribution of the observations is Gaussian). Taking the empirical counterparts of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ leads to

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \quad \text{and} \quad \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t.$$

These estimators are non robust. One single observation among n can break the empirical mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{S}$, which gives *empirical* breakdown points equal to $1/n$, and *asymptotic* breakdown points (obtained for $n \longrightarrow \infty$) equal to zero. The influence functions are not bounded either.

Outliers in higher dimensions

Multivariate analysis

Trimmed estimators (MCD), M-estimators, S-estimators etc. are available in stata using the command `robmv`. We focus on the projection-based [Stahel, 1981] and [Donoho, 1982] estimator.

A multivariate outlier is a point that lies far away from the bulk of data in any direction and can be seen as an outlier in some univariate projection.

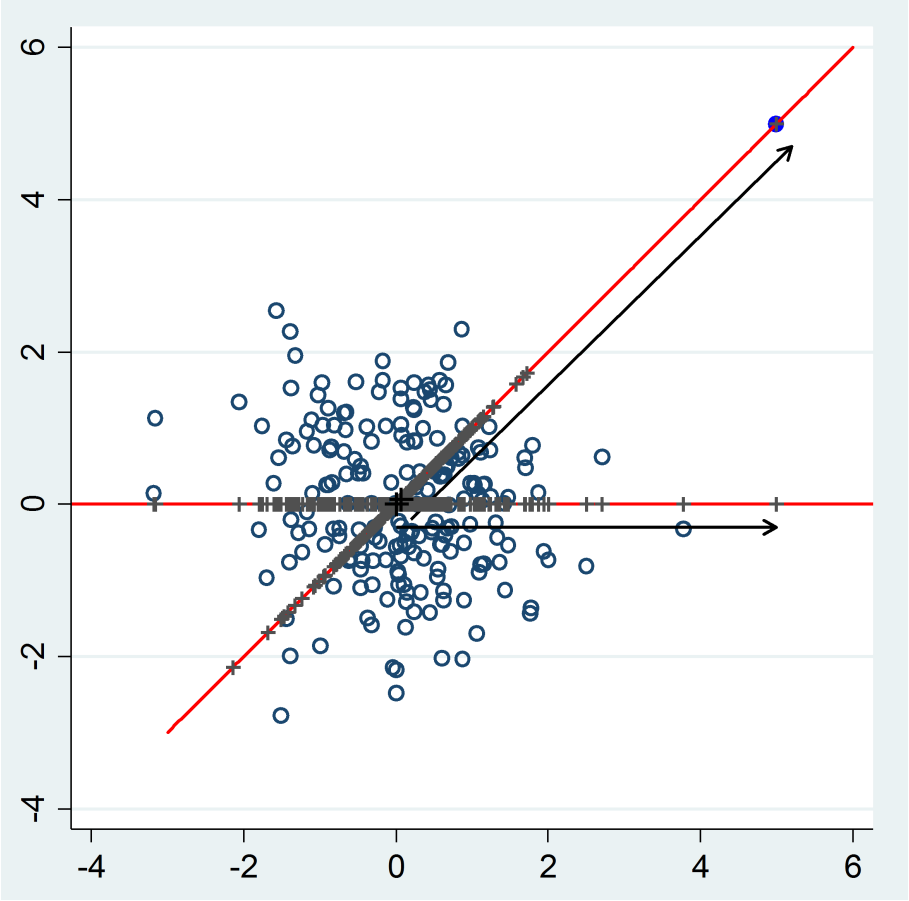
Definition

Given a direction $\mathbf{a} \in \mathbb{R}^p$ with $\|\mathbf{a}\| = 1$, denote by $\mathcal{X}_{\mathbf{a}}^{(n)} = \{\mathbf{x}_1^t\mathbf{a}, \ldots, \mathbf{x}_n^t\mathbf{a}\}$ the projection of the dataset along $\mathbf{a}$. Let $\widehat{\mu}$ and $\widehat{\sigma}$ be robust univariate location and dispersion statistics. The outlyingness of a point along $\mathbf{a}$ is defined as $r_{\mathbf{a}}(\mathbf{x}) = \frac{\left|\mathbf{x}^t\mathbf{a} - \widehat{\mu}\left(\mathcal{X}_{\mathbf{a}}^{(n)}\right)\right|}{\widehat{\sigma}\left(\mathcal{X}_{\mathbf{a}}^{(n)}\right)}$ and the *Stahel-Donoho outlyingness* of $\mathbf{x}$ as $r(\mathbf{x}) = \sup_{\mathbf{a}\in\mathcal{S}_p} r_{\mathbf{a}}\left(\mathbf{x}; \mathcal{X}^{(n)}\right)$ with $\mathcal{S}_p = \{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\| = 1\}$.

Outliers in higher dimensions

Stahel-Donoho estimator

If the data are normally distributed, the global outlyingness measures SDO_i are asymptotically χ^2_p distributed [Maronna and Yohai, 1995]



Outliers in higher dimensions

Modified Stahel-Donoho estimator

The Stahel-Donoho method is only suited for elliptical data as it assumes that the scale on the lower and upper sides of the median to be equal.

The Stahel-Donoho outlyingness measure can be modified to take this into account.

The *asymmetrical* outlyingness with respect to $\mathcal{X}^{(n)}$ of a point $\mathbf{x} \in \mathbb{R}^p$ along $\mathbf{a}$ can be defined as

$$\mathrm{ASO}_{\mathbf{a}}\left(\mathbf{x}; \mathcal{X}^{(n)}\right) = \begin{cases} \dfrac{\mathbf{x}^t\mathbf{a} - Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)})}{2c\left[Q_{0.75}(\mathcal{X}_{\mathbf{a}}^{(n)}) - Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)})\right]} & \text{if } \mathbf{x}^t\mathbf{a} \geq Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)}) \\ \dfrac{Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)}) - \mathbf{x}^t\mathbf{a}}{2c\left[Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)}) - Q_{0.25}(\mathcal{X}_{\mathbf{a}}^{(n)})\right]} & \text{if } \mathbf{x}^t\mathbf{a} < Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)}) \end{cases}$$

where Q_η is η^{th} percentile the of the projected dataset $\mathcal{X}_{\mathbf{a}}^{(n)}$ and $c = 0.7413$.

The *asymmetrical* (global) *outlyingness of* $\mathbf{x}$ *with respect to* $\mathcal{X}^{(n)}$ is then given by $\mathrm{ASO}\left(\mathbf{x}; \mathcal{X}^{(n)}\right) = \sup_{\mathbf{a} \in \widehat{\mathcal{S}}_p} \mathrm{ASO}_{\mathbf{a}}\left(\mathbf{x}; \mathcal{X}^{(n)}\right)$

Outliers in higher dimensions

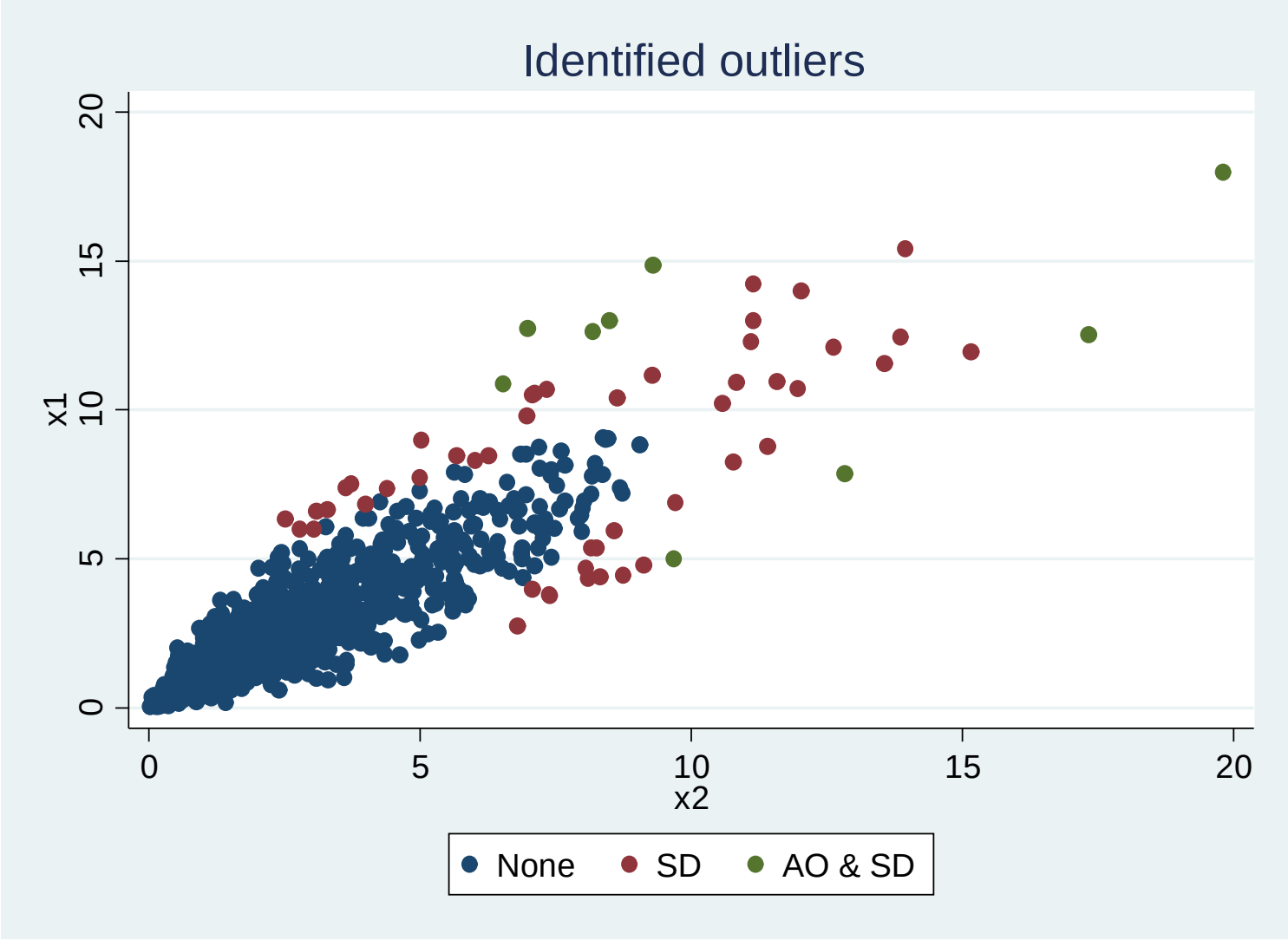
Modified Stahel-Donoho estimator

Outliers can be identified using the generalized boxplot of ASO_i distances as described before.

```
clear
graph drop _all
set obs 1000
matrix C=(1,0.9\0.9,1)
drawnorm x1 x2, corr(C)
replace x1=invchi2(3,normal(x1))
replace x2=invchi2(3,normal(x2))
sd x1 x2, gen(a0 b0) level(0.99)
sdasym x1 x2, generate(a b) level(0.99)
twoway (scatter x1 x2 if a0==0&a==0) (scatter x1 x2 if
a0==1&a==0) (scatter x1 x2 if a0==1&a==1), legend(order( 1
"None" 2 "SD" 3 "AO & SD") rows(1)) title("Identified
outliers")
```

Outliers in higher dimensions

Modified Stahel-Donoho estimator



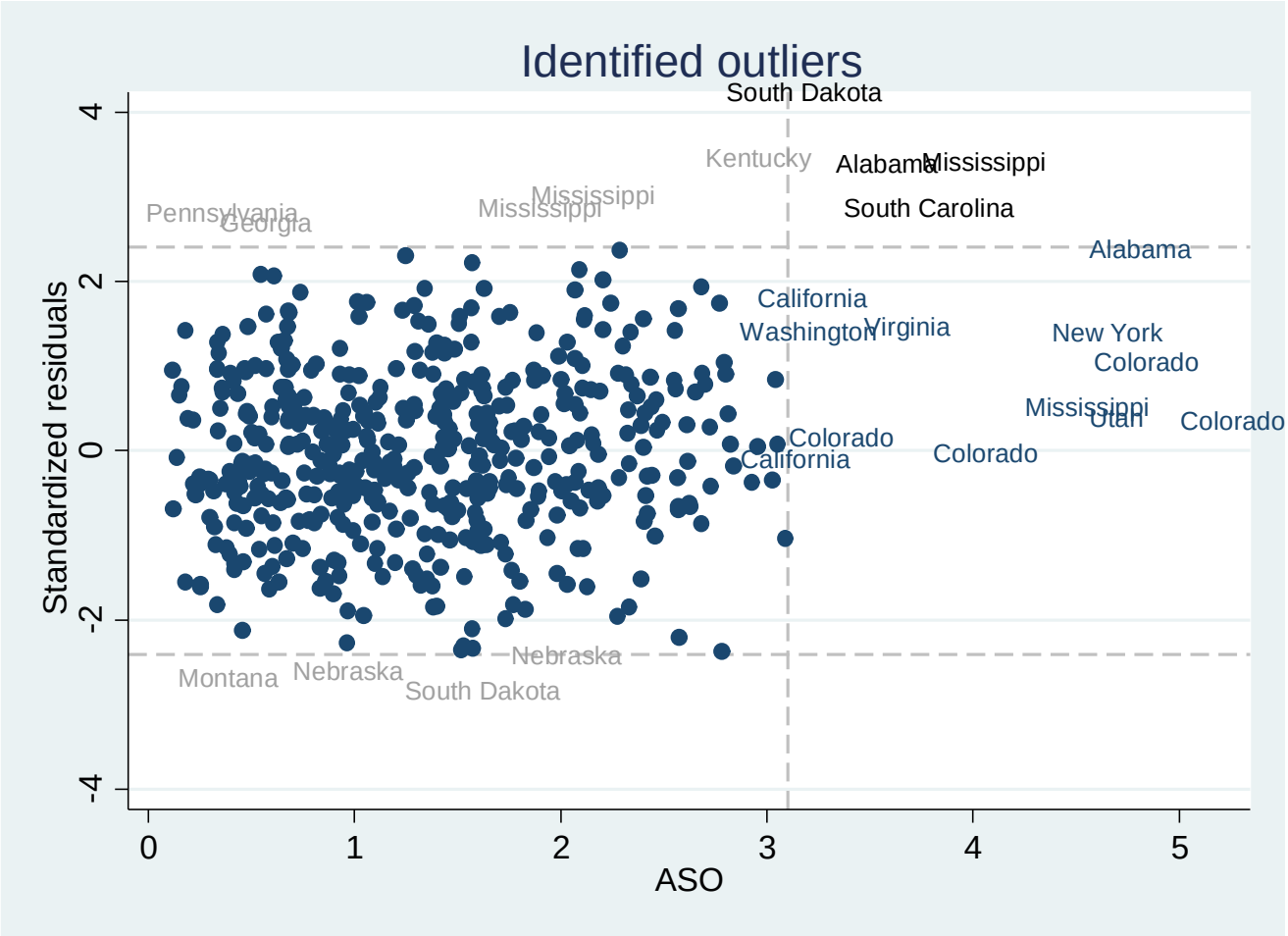
Unmasking the outliers

Obesity prevalence in a sample of US counties

```
use ...\Diabetes.dta

robreg mm perdiabet percphys perco, eff(60) hausman

predict res

replace res=(perdiabet-res)/e(scale)

sdasym percphys percob, gen(a b) level(0.99)

local l=invnorm(0.99)

local u=e(cutoff)
scatter res b, xline(‘u’) yline( ‘l’ -‘l’)
```

Unmasking the outliers

Obesity prevalence in a sample of US counties
[Rousseeuw and van Zomeren, 1990]



Binary models

"There are 10 kinds of people in the world: those who understand binary numerals, and those who don't."

Ian Stewart

Standard Logit

Let y be a 1-0 variable indicating the realization of a specific event. We wish to predict this outcome by means of a linear combination of different explanatory variables $x_1, x_2, \ldots, x_p$.

Let us use the notation $p_i(\boldsymbol{\beta}) = F\left(\boldsymbol{\beta}^t \mathbf{x}_i\right) = \mathrm{P}\left(y_i = 1 | \mathbf{x}_i\right)$. The maximum (log)likelihood estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$ of $\boldsymbol{\beta}$ is then defined as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \ln \left[\prod_{i=1}^{n} F\left(\boldsymbol{\beta}^t \mathbf{x}_i\right)^{y_i} \left(1 - F\left(\boldsymbol{\beta}^t \mathbf{x}_i\right)\right)^{1-y_i} \right]$$

$$= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left[-y_i \ln F\left(\boldsymbol{\beta}^t \mathbf{x}_i\right) - (1 - y_i) \ln \left(1 - F\left(\boldsymbol{\beta}^t \mathbf{x}_i\right)\right) \right]$$

$$= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} d_i^2(\boldsymbol{\beta}) \text{ where } d_i(\boldsymbol{\beta}), \text{ deviance residuals, is}$$

$$\sqrt{-y_i \ln F\left(\boldsymbol{\beta}^t \mathbf{x}_i\right) - (1 - y_i) \ln \left(1 - F\left(\boldsymbol{\beta}^t \mathbf{x}_i\right)\right)} \operatorname{sign}\left(y_i - F\left(\boldsymbol{\beta}^t \mathbf{x}_i\right)\right)$$

Robust Logit

[Pregibon, 1982] proposed to robustify the method using a similar logic to M-estimators. He suggests to find $\widehat{\boldsymbol{\beta}}_{\mathrm{P}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \lambda\left(d_i^2\left(\boldsymbol{\beta}\right)\right)$ where $\lambda\left(v\right) = \begin{cases} v & \text{if } v \leq q \\ 2\sqrt{qv} - q & \text{if } v > q \end{cases}$

[Bianco and Yohai, 1996] considered a *bounded* function γ instead of λ introducing a correction to ensure the Fisher-consistency: $\widehat{\boldsymbol{\beta}}_{\mathrm{BY}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \left[\gamma\left(d_i^2\left(\boldsymbol{\beta}\right)\right) + C\left(\boldsymbol{\beta}\right)\right]$.

[Croux and Haesbroeck, 2003] suggest to choose γ such that $\gamma'\left(v\right) = \begin{cases} \exp\left(-\sqrt{0.5}\right) & \text{if } v \leq 0.5 \\ \exp\left(-\sqrt{v}\right) & \text{if } v > 0.5 \end{cases}$ and weight the obervations using

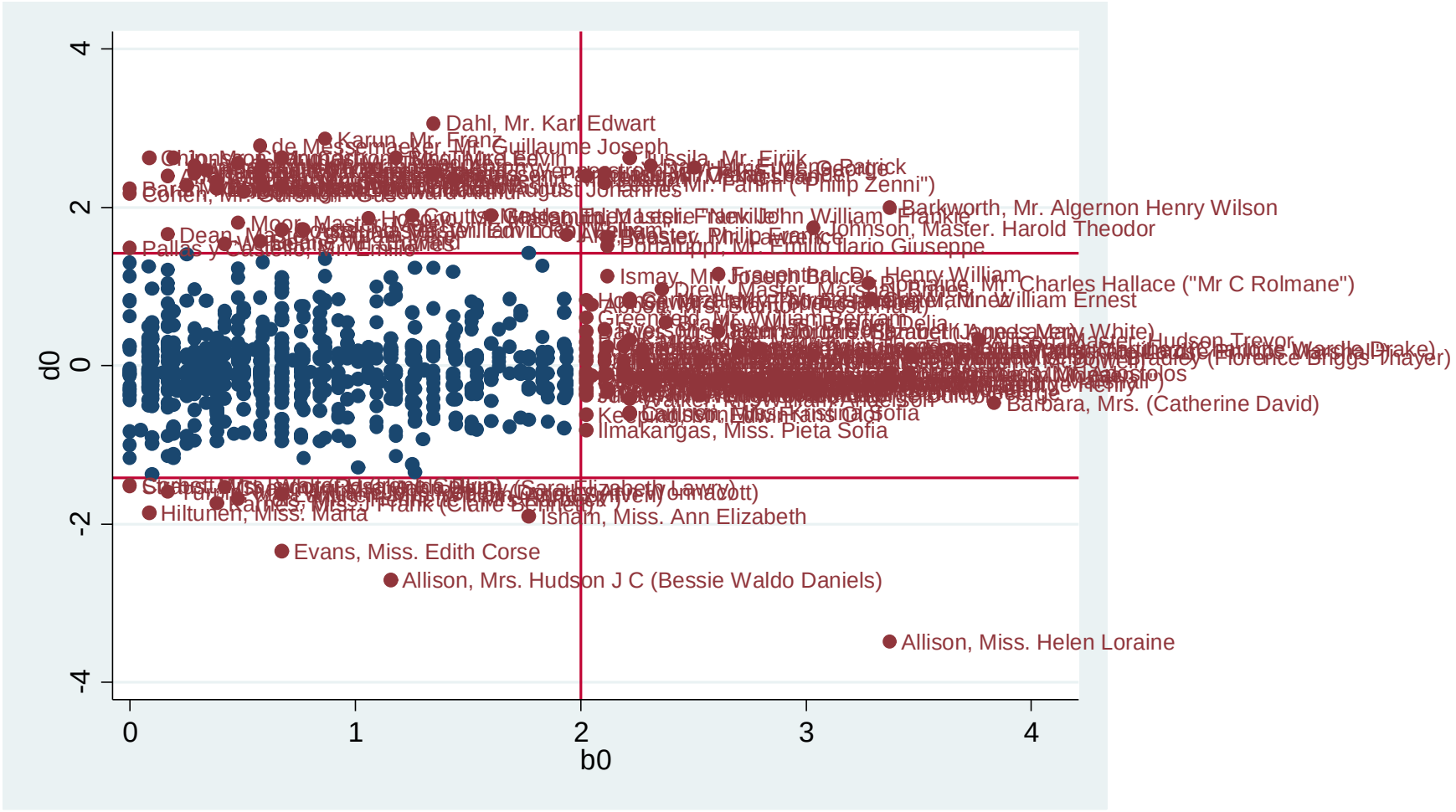
$w_i = \begin{cases} 1 & \text{if } d^2\left(\underline{\mathbf{x}}_i, \widehat{\boldsymbol{\mu}}_n; \widehat{\boldsymbol{\Sigma}}_n\right) \leq \chi^2_{p;0.975} \\ 0 & \text{else;} \end{cases}$ to reduce leverage effect.

Outliers in the Titanic accident

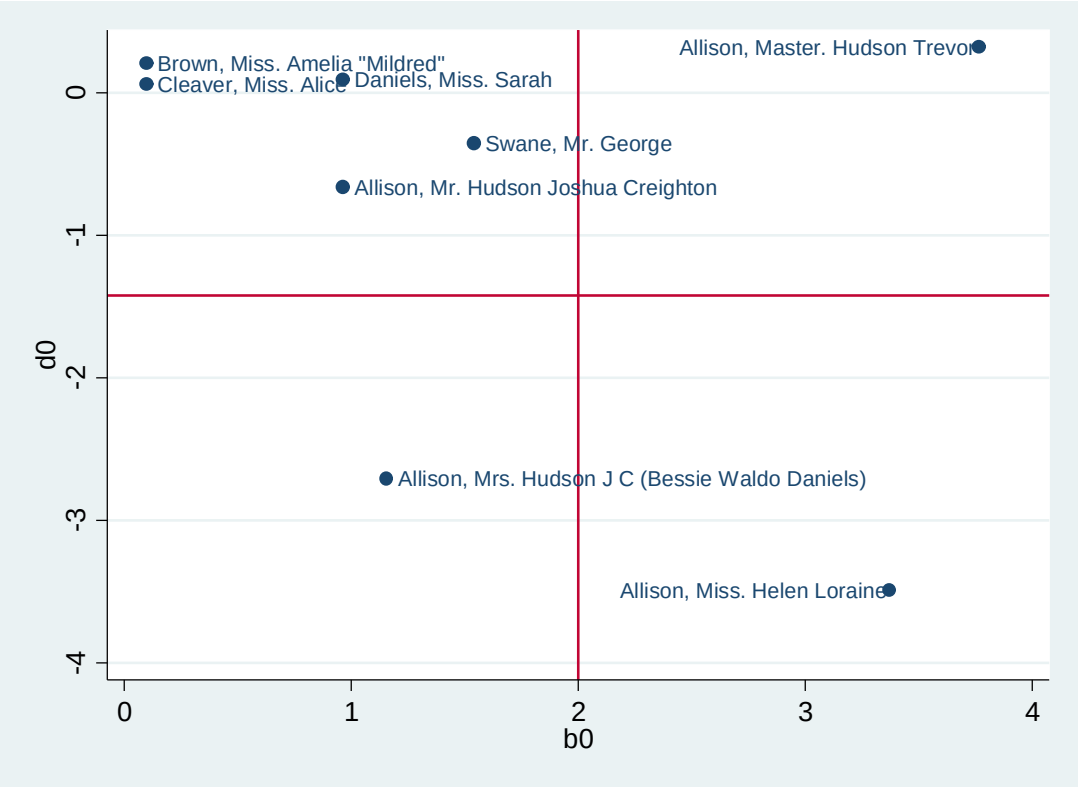
Identify the outliers in the Titanic accident

- Run a robust Logit of survival (y) on age, gender and class dummy
- The stata command is `roblogit`
- Estimate deviances defined as $d_i = -y_i ln(\hat{p}) - (1-y) ln(1-\hat{p})$
- The distribution of deviances is unknown, use the propose technique to identify outliers
- Use the outlier identification tool of [Rousseeuw and van Zomeren, 1990]
- Zoom on interesting outliers

Outliers in the Titanic accident



Titanic accident - Ticket No. 113781, cabins C22-C26



Mr H.J.C. Allison (30)†	Head
Mrs H.J.C. Allison (25)†	Spouse
Miss H.L. Allison (2)†	Daughter
Master H.T. Allison (1)	Son
Miss A.C. Cleaver (22)	Nurse
Miss S. Daniels (33)	Nurse
Miss A.M. Brown (18)	Cook
Mr. G. Swane (26)†	Chauffeur

Conclusion

Several robust estimators exist in stata

- Descriptive statistics (trimmed estimators, quantile based and pairwise based)
  - **Location**
  - **Scale**
  - **Skewness**
  - Tail heavyness
- Regression analysis
  - Linear regression (**M-estimators**, Generalized M-estimators, **S-estimators**, **MM-estimators**)
  - Instrumental variables
  - **Logit**
  - Panel data
- Multivariate analysis
  - M-estimators, S-estimators, MCD estimator, MVE estimator, **Projection based estimators**, Robust Principal Component

References I

- Bianco, A. M. and Yohai, V. J. (1996).
  Robust estimation in the logistic regression model.
  In Rieder, H., editor, Robust Statistics, Data Analysis, and Computer Intensive Methods. In Honor of Peter Huber's 60th Birthday, pages 17–34. Springer, New York.
- Bruffaerts, C., Verardi, V., and Vermandele, C. (2014).
  A generalized boxplot for skewed and heavy-tailed distributions.
  Statistics & Probability Letters, 95(C):110–117.
- Brys, G., Hubert, M., and Struyf, A. (2004).
  A robust measure of skewness.
  Journal of Computational and Graphical Statistics, 13(4):996–1017.

References II

- Croux, C., Dhaene, G., and Hoorelbeke, D. (2003).
  Robust standard errors for robust estimators.
  Discussions Paper Series (DPS) 03.16, Center for Economic Studies, KULeuven.
- Croux, C. and Haesbroeck, G. (2003).
  Implementing the bianco and yohai estimator for logistic regression.
  Computational Statistics and Data Analysis, 44(1-2):273–295.
- Dehon, C., Gassner, M., and Verardi, V. (2012).
  Extending the hausman test to check for the presence of outliers.
  Advances in Econometrics, 29 - Essays in Honor of Jerry Hausman:435–453.
- Donoho, D. (1982).
  Breakdown Properties of Multivariate Location Estimators.
  PhD thesis, Harvard University.

References III

- Edgeworth, F. Y. (1887).
  On observations relating to several quantities.
  Hermathena, 6:279–285.
- Hansen, L. P. (1982).
  Large sample properties of generalized method of moments estimators.
  Econometrica, 50:1029–1054.
- Hodges, Jr., J. L. and Lehmann, E. L. (1963).
  Estimates of location based on rank tests.
  Annals of Mathematical Statistics, 34(2):598–611.
- Huber, P. J. (1981).
  Robust Statistics.
  John Wiley & Sons, New York.

References IV

- Maronna, R. A., Martin, D. R., and Yohai, V. J. (2006).
  Robust Statistics. Theory and Methods.
  John Wiley & Sons, Chichester.
- Maronna, R. A. and Yohai, V. J. (1995).
  The behavior of the Stahel-Donoho robust multivariate estimator.
  Journal of the American Statistical Association, 90(429):330–341.
- Pregibon, D. (1982).
  Resistant fits for some commonly used logistic models with medical applications.
  Biometrics, 38:485–498.
- Rousseeuw, P. J. and Croux, C. (1993).
  Alternatives to the median absolute deviation.
  Journal of the American Statistical Association, 88(424):1273–1283.

References V

- Rousseeuw, P. J. and Leroy, A. M. (1987).
  Robust Regression and Outlier Detection.
  John Wiley & Sons, New York.
- Rousseeuw, P. J. and van Zomeren, B. C. (1990).
  Unmasking multivariate outliers and leverage points.
  Journal of the American Statistical Association, 85(411):633–639.
- Stahel, W. A. (1981).
  Breakdown of covariance estimators.
  In Research Report 31. Fachgruppe fur Statistik, E.T.H. Zurich, Switzerland.
- Yohai, V. J. (1987).
  High breakdown-point and high efficiency robust estimates for regression.
  The Annals of Statistics, 15(2):642–656.

Procedure

Transformation

1. Center and reduce the data: $x_i^* = \frac{x_i - Q_{0.5}(\{x_j\})}{\text{IQR}(\{x_j\})}$
2. Shift the dataset to obtain only strictly positive values: $r_i = x_i^* - \min(\{x_j^*\}) + 0.1$
3. Standardize r_i to map x_i on $(0,1)$: $\widetilde{r}_i = \frac{r_i}{\min(\{r_j\}) + \max(\{r_j\})}$
4. Consider the inverse normal transformation $w_i = \Phi^{-1}(\widetilde{r}_i)$
5. Center and reduce the values w_i: $w_i^* = \frac{w_i - Q_{0.5}(\{w_j\})}{\zeta \text{IQR}(\{w_j\})/1.3426}$
6. Adjust the distribution of the values w_i^* $(i = 1, \ldots, n)$ by the Tukey $T_{\widehat{g}^*, \widehat{h}^*}$ distribution:

$$\widehat{g} = \frac{1}{z_{0.9}} \ln\left(-\frac{P_{0.9}\left(\{w_j^*\}\right)}{P_{0.1}\left(\{w_j^*\}\right)}\right), \qquad \widehat{h} = \frac{2\ln\left(-\widehat{g}\,\frac{P_{0.9}\left(\{w_j^*\}\right)P_{0.1}\left(\{w_j^*\}\right)}{P_{0.9}\left(\{w_j^*\}\right)+P_{0.1}\left(\{w_j^*\}\right)}\right)}{z_{0.9}^2}$$

7. Select the rejection bounds (L_-^*, L_+^*) using specific quantiles of the adjusted distribution (here $P_{0.35}$ and $P_{99.65}$) and do the inverse transformation